**The Dissertation Committee for Amy Broward Weisenburgh Certifies that this is the approved version of the following dissertation:**

# Developing a Screening Measure for At-Risk and Advanced Beginning Readers to Enhance Response-To-Intervention Frameworks Using the Rasch Model

**Committee:**

Sharon Vaughn, Supervisor

Barbara Dodd, Co-Supervisor

Sylvia Linan-Thompson

Mark O'Reilly

Herb Rieth

# Developing a Screening Measure for At-Risk and Advanced Beginning Readers to Enhance Response-To-Intervention Frameworks Using the Rasch Model

*by*

**Amy Broward Weisenburgh, B.S., M.ED.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

The University of Texas at Austin

**December 2011**

Dedicated to my parents

Louis and Mary Weisenburgh

# Acknowledgements

There are several individuals I wish to thank for supporting my doctoral studies. I am particularly thankful to my advisor Sharon Vaughn for developing my understanding of group experimental design principles and best practices in intervention research. As my dissertation co-chair, Sharon encouraged me to step outside of my comfort zone and pursue my emerging interests in applying item response theory models to the assessment of exceptional readers. I would like to thank my other co-chair Barbara Dodd for igniting my fascination with latent trait models while always being available to answer seemingly endless questions despite large piles of work always accumulating on her desk. Together, Sharon and Barbara encouraged me to stretch beyond my limits. Their instruction, continuous guidance, and intellectual support have made a profound impact on my interests and the direction of my future research pursuits.

I wish to thank Sylvia Linan-Thompson for her advice and support in extending my studies to the assessment of English Language Learners in developing countries. In addition, I am thankful for the friendship that has emerged. I am also thankful for the invaluable support, encouragement, and friendship of Joanne Robbins and Joe Layng throughout my graduate studies.

I am grateful to Herb Rieth and Mark O'Reilly for lending a behavior analytic perspective to my committee. I would also like to acknowledge the contribution of Lynn Chen to the operationalization and analysis of this research.

Throughout my studies at Texas, I have had the privilege to learn from Don Hammill. Don has provided me with tremendous insight into innovative test development strategies. He also served as a constant sounding board that challenged me to apply my understanding of classical test theory and item response theory techniques to practical issues of test development.

I have dedicated my dissertation to my father and mother, Louis and Mary Weisenburgh. My parents instilled in me a passion for learning in a loving and supportive home where trying my best was always more important than the actual outcome. Thank you Mom and Dad - words could never properly express how much I appreciate the endless ways you have supported my development as a student and as a person.

This journey could not have been completed without the support, friendship, and love of my amazing husband and best friend. Quin, thank you for your encouragement and patience as well as the sacrifices that were made that allowed me to complete my studies.

Finally, I would like to thank the educators and students in the townships of Port Elizabeth, South Africa. Were it not for your inspiration and belief in education as a tool that can change the world, I would have given up on this course long ago.

*It is good to have an end to journey towards;*

*but it is the journey that matters in the end.*

- Urusla K. Le Guin

# Developing a Screening Measure for At-Risk and Advanced Beginning Readers to Enhance Response-To-Intervention Frameworks Using the Rasch Model

Publication No._____

Amy Broward Weisenburgh, Ph.D.

The University of Texas at Austin

December, 2011

Supervisors: Sharon Vaughn & Barbara Dodd

The Rasch model was employed to analyze the psychometric properties of a diagnostic reading assessment and then create five short forms (n = 10, 16, 22, 28, 34 items) with an optimal test information function. The goal was to develop a universal screening measure that second grade teachers can use to identify advanced and at-risk readers. These groups were targeted because both will need differentiated instruction in order to improve reading skills. The normative dataset of a national reading test developed with classical test theory methods was used to estimate person and item parameters. The measurement precision and classification accuracy of each short form was evaluated with the second grade students in the normative sample. Compared with full bank scores, all short forms produced highly correlated scores. The degree to which each short form

identified exceptional readers was also analyzed. In consideration of classification accuracy and time-efficiency, the findings were most robust for the 10-item form.

## *Table of Contents*

## *List of Tables*

## *List of Figures*

# CHAPTER I: INTRODUCTION

Reading is an essential skill required to access and understand nearly all subjects in school. Higher individual and family reading levels are positively correlated with higher income levels, which in turn are positively associated with quality of life indicators such as health and life expectancy (Barton & Jenkins, 1995). Problematically, many students in the United States have difficulties learning to read. For example, according to recent estimates, 67% of fourth grade students lack proficient reading skills (National Center for Educational Statistics, 2010). Some of these students have or will be diagnosed with a learning disability (LD). Approximately 5% of the school aged population is currently identified as learning disabled while nearly half of these students were qualified for special education services due to difficulties learning to read (President's Commission on Excellence in Special Education, 2002). If these students are not identified and the problem is not remediated before the end of second grade, it is unlikely they will ever catch up to their peers (Lyon et al., 2001), which may result in long-term social problems (See Gellert & Elbro, 1999 for review).

## *Importance of Early Identification*

Research suggests that early identification and intervention is the best way to reduce the number of students later identified with a learning disability in the area of reading (Snow, Burns, & Griffin, 1998). Some experts further contend that the number of children typically identified as poor readers and served through either special education or compensatory education programs could be reduced by up to 70% given effective early identification and prevention measures (Lyon et al., 2001). With the recent reauthorization of the Individuals with Disabilities Act (IDEA, 2004), states now have the option of using

an alternative framework referred to as Response To Intervention (RtI) to address federal screening mandates and diagnose students with learning disabilities.

## *RtI as Identification and Instructional Delivery Framework*

RtI is often presented in the context of priorities underscored by the No Child Left Behind legislation which includes an emphasis on screening all students for reading difficulties in the early school years; placement in early intervention programs; and careful monitoring of progress with accountability for results (Jenkins, Hudson, & Johnson, 2007). RtI is most often conceptualized as a three-tiered model in which instruction at each successive tier is more intense and explicit while group size is reduced (Vaughn & Linan-Thompson, 2003). As summarized by Johnson, Jenkins, Petscher, and Catts (2009),

> RtI is a multitiered instructional and service delivery model designed to improve student learning by providing high-quality instruction, intervening early with students at-risk for academic difficulty, allocating instructional resources according to students' needs, and distinguishing between students whose reading difficulties stem from experiential and instructional deficits as opposed to a learning disability. Derived from the prevention sciences, RtI represents an attempt to identify and help struggling readers early before academic problems develop into intractable deficits (p.174).

While the RtI educational model was originally designed to identify and provide specialized educational services to struggling readers, students with advanced academic skills could also benefit from efforts to match instruction to instructional need (Hong & Hong, 2009; Wright, Horn, & Sanders, 1997). As noted by Brown and Abernathy (2009), in practice, RtI is about effective teaching practices which includes preassessing students through a strategic process, making instructional modifications in accordance with instructional needs, and monitoring student progress employing a tiered approach so

higher student outcomes can be realized. RtI therefore has important implications for the education of highly capable students as a model for policy development.

In alignment with federal mandates, screening measures used within RtI frameworks target students at-risk for academic failure. Though no federal laws currently protect the legal rights of highly capable students, approximately 30 states do have a mandate to service "gifted" children (Council of State Directors of Programs for the Gifted, 1994). Most school districts require students to be referred or nominated before being formally assessed for advanced educational programs. Problematically, research indicates that the referral process is often biased against African American, Hispanic, and economically disadvantaged students (McBee, 2006). These students could greatly benefit from the universal screening process inherent to RtI frameworks. This dissertation addresses the needs of academically advanced and at-risk students by developing a screening measure that teachers can use to simultaneously identify both groups of readers. In the next section I discuss issues surrounding universal screening measures in beginning reading as wells as methods of improving the accuracy of Tier 1 screening decisions.

## *Universal Screening Measures in Beginning Reading*

Given the importance of identifying and remediating beginning reading problems, the No Child Left Behind federal legislation (NCLB, 2001) specifically stipulates "the use of rigorous diagnostic and screening assessment tools" (Title I, Part B, Sec. 1201: Reading First). Local Education Agencies are therefore required to use screening, diagnostic, progress monitoring, and outcome testing instruments with appropriate reliability and validity to facilitate the identification and education of students at-risk for or already identified with reading disabilities.

The defining feature of a screening measure is its ability to accurately classify students as at-risk or not at-risk for failure. The accuracy of a screening measure is most often characterized by its degree of sensitivity and specificity. Sensitivity refers to the accuracy of the screener to identify students that are at-risk. Specificity refers to the accuracy of the screener to identify students that are not at-risk (Jenkins, Hudson, & Johnson, 2007). The acceptable levels of sensitivity and specificity vary by field and according to the intended purpose of the test. If the purpose of the test is to ensure that truly at-risk students for failure in reading are identified while not wasting limited human and financial resources, some suggest a rigorous sensitivity standard of 0.90 and a specificity standard close to 0.90 (Jenkins, 2003).

Despite federal mandates requiring the use of high quality measures, emerging evidence suggests that many widely-used universal screening tools in beginning reading are technically inadequate. For instance, Jenkins, Hudson, and Johnson (2007) reviewed the classification accuracy of reading screeners used in elementary schools as reported in studies published since 1998 (n = 11). The average sensitivity and specificity levels for popular measures used in Kindergarten (73%, 82%), Grade 1 (62%, 86%), Grade 2 (49%, 86%), and Grade 3-4 (80%, 84%), respectively, were below recommended guidelines. As a result, many experts agree that too many beginning readers are either being either unidentified or misidentified (Johnson, Jenkins, Petscher, & Catts, 2009; Reidel, 2007). To the extent identifying "at-risk" students is the first step in remediating academic deficits and preventing long-term failure, there is a tremendous need for more precise measures. In consideration of these results, and given the importance of identifying struggling readers before the end of second grade, this study focuses on developing a screening measure for second grade students.

## *Methods of Improving the Accuracy of Tier 1 Screening Decisions*

There are several ways to improve the accuracy of Tier 1 screening decisions. These include (1) using expanded screens that measure more than one skill, (2) adjusting the cut score used to define the outcome, and (3) improving the quality of the measure with modern test theory techniques. Each of these methods will be explained in more detail below.

*Expanded screens*. In an attempt to improve Tier 1 screening decisions, researchers have analyzed the accuracy of screens based on a single measure compared to screens combining more than one measure. Foorman et al. (1998) and O'Connor and Jenkins (1999) reported improved accuracy in identifying at-risk readers by combining scores on several measures. One limitation of this approach is the cost of administering additional measures in terms of time and personnel (Jenkins, Hudson, & Johnson, 2007).

The research conducted by Foorman et al. (1998) and O'Connor and Jenkins (1999) further revealed that screens that are valid for one grade level might not be valid for another. This means that in order to adequately detect differences in individual reading development, screens should be sensitive to developmental reading skills across grade levels. In kindergarten, the greatest growth occurs in phonemic awareness, letter and sound knowledge, and vocabulary. By first grade, students continue to develop phonemic awareness, letter and sound knowledge, and vocabulary, however the greatest growth occurs in phonemic spelling, decoding, word identification, and text reading. In second and third grade, reading growth is reflected in the number and type of words students can read, the difficulty of texts they can read and comprehend, and the fluency with which these tasks are accomplished. Beyond third grade, comprehension of more difficult texts becomes the primary measure of reading development (Johnson, Jenkins, & Hudson, 2007). Given these findings, the present investigation focuses on the identification of at-

risk and advanced second grade readers using a single measure that requires both word identification skills and vocabulary skills.

*Adjusting cut scores*. The accuracy of Tier 1 screening decisions can also be improved by changing the cut score used to define the outcomes. A cut score represents the dividing line between those that are not at-risk and those that are potentially at-risk. The technique of adjusting cut scores involves changing the cut points distinguishing between "at-risk" and "not at-risk" students to improve classification accuracy. However, in practice, decreasing the cut score needed to qualify as not at-risk often increases the accuracy of identifying truly at-risk students (i.e., higher sensitivity), but decreases the accuracy of identifying those that are truly not at-risk (i.e., lower specificity). The result is over-referrals. Alternatively, increasing the cut point needed to qualify as not at-risk, often increases the accuracy of identifying not at-risk students (i.e., higher specificity), but decreases the accuracy of identifying those truly at-risk (i.e., lower sensitivity). The result is under-referrals.

*Modern test theory models*. The accuracy of Tier 1 screening decisions can also be increased by using modern test theory models to improve the structure of the measurement scale (Schultz-Larsen, Kreiner, & Lomholt, 2007). This dissertation is designed to explore and analyze the effects of using the one-parameter item response theory model, referred to as the Rasch Model, to develop a screening measure to identify at-risk and advanced second grade readers. The theoretical framework of item response theory models in general, and the Rasch model in particular, will be described below.

## *Item Response Theory Models*

### Theoretical Framework

One of the critical elements of scientific measurement is comparability. Measurement instruments used in the physical sciences must be either equivalent or equatable. They should lead to comparisons that are the same irrespective of qualitative factors. A person's height is the same regardless of the color, material, or unit (e.g., inches, feet, or meters) scale of the ruler. However, if my ruler at home says I am 6 feet tall, but the one at the hospital says I am only 5 feet tall – most would recognize a problem and question the value of at least one of the rulers. Conversely, when it is reported that a student weighs 100 pounds, no one demands to see the scale or know when and where it was made. Rather it is assumed that the estimate is completely independent of the circumstances surrounding its construction. In other words, it is assumed to be an "objective" or "invariant" measure. Historically, this logic of measurement in the physical sciences has not been extended to measurement in the cognitive sciences.

The majority of educational tests developed since the turn of the century are based on an approach to measurement known as classical test theory. In classical test theory, the notion of ability is expressed as the *true score*, which is defined as the expected value of observed performance on the test of interest. Ability estimates are calculated as the total number of correct responses to a given set of test items. Within this framework, if a test is comprised of items that are difficult for a student, then s/he will appear to have a low ability. Alternatively, if a test is comprised of items that are easy for a student, then s/he will appear to have a higher ability. In this case, two rulers made of different material will yield two different ability estimates for the same person. This mental measurement conundrum is further confounded by the fact that the difficulty of an item is quantified according to the proportion of examinees that answer the item correctly. Thus, the ability

estimates of examinees depend on the accuracy of performance on items that are classified as easy or difficult, yet whether an item is classified as easy or difficult depends on the abilities of examinees in the normative sample. Given the degree to which the nature of the population used to develop an assessment influences ability estimates, item statistics, and test statistics, classical test theory developers often invest significant time and financial resources to ensure representative standardization samples.

Beginning in the early 1950's, alternative measurement models were developed to address the issues of dependency between measurement tools and the individuals being measured. Lord (1952), Birnbaum (1958), and Rasch (1960) pioneered a new approach to test development based on mathematical models for dichotomously-scored items that exhibited the property of independence or "specific objectivity." Specific objectivity means that any subset of items that measure the same trait can be used to estimate an individual's ability on the same scale of measurement and the resulting item parameter estimates will be invariant across different samples of individuals used to calibrate the items. Lord and Novick (1968) were among the first to provide a comprehensive description of modern test theory which they called item response theory (IRT). Item response theory refers to a technique of modeling the mathematical relationship between the latent variable (often called "trait" or "ability") underlying performance and scores on individual items and groups of items. As such, item response theory is also often referred to as latent trait theory. The latent trait variable is conceptualized as a continuous, unidimensional construct that attempts to explain the covariance among item responses (Steinberg & Thissen, 1995).

Item response theory models can be used to overcome many of the limitations inherent in tests designed with classical test theory methods. For instance, test developers can specify how well *individual items* discriminate between individuals who do or do not

possess a particular skill at various ability levels. Such specification makes it possible to examine the expected degree of measurement error on an item-by-item basis at different ability levels rather than relying on a single gross standard error of measurement applied equally to all ability levels - as required with classical approaches. In this way, modern models can be used to quantify how the addition or deletion of a single item will impact measurement precision and error at each ability level prior to administration. Therefore, test items can be selected to maximize information and minimize error at targeted ability levels. Through this process, shorter forms of comprehensive tests can be developed with reliability equal to or greater than their conventional counterparts (Dodd, 1984). These forms can be used to more quickly and accurately identify learners performing significantly behind or beyond conventional standards. This issue is particularly acute for students with unique learning needs for whom test scores can dictate the provision or denial of specialized education services – which may or may not be needed.

## THE RASCH MODEL

Of all item response theory models proposed for person measurement, the Rasch model (often referred to as the one-parameter logistic model) is considered to be the simplest and most efficient (Bond & Fox, 2007; Embretson & Reise, 2000). Similar to other item response theory models, the Rasch model transforms ordinal data to an interval scale of scores for item difficulties and person abilities. However the Rasch model is unique in that scores are reported in units called *logits*. Similar to the ruler that measures length in inches, the logit scale measures item difficulty on one side of the ruler and person ability on the other side of the ruler in logits. Given item difficulty and person ability estimates on the same interval scale, it is possible to calculate how much more difficult one item is compared to another or how much more proficient one student is compared to another. Just like two inches is twice as long as one inch, an item with two

logits is twice as difficult as an item with one logit. Similarly, a person with a logit score of 15 has three times the ability of a person with a logit score of five. The Rasch model is therefore probabilistic such that the distance between an item's difficulty and a person's ability governs the probability of being successful on any given test item. The basic premise is that there is a higher probability of correctly answering easier items than correctly answering more difficult items. In turn, given harder questions, there is a higher probability that more proficient individuals will answer correctly and a lower probability that less proficient individuals will answer correctly (Rasch, 1980). The distinguishing feature of the Rasch model compared to other item response theory models is that the raw score is a sufficient statistic. This means that student ability can be readily estimated and more easily interpreted by practitioners in consonance with standard practice. Conversely, when the two- or three-parameter models are used, specialized software is required to calculate ability estimates and more advanced training is needed to ensure outcomes are interpreted accurately (Embretson & Reise, 2000).

Despite the many advantages of using item response theory models in general and the Rasch model in particular to develop and refine assessments for exceptional learners, there have been surprisingly few studies published in the field of special education. An extensive literature search revealed only one published study that applied item response theory techniques to develop a screening measure for beginning readers. Foorman and colleagues (1998) applied the two-parameter model to develop a screening measure by selecting items with difficulty parameters nearest to the established cut points. However, this study did not (1) analyze the extent to which test length impacts measurement precision and decision accuracy or (2) attempt to design a measure to simultaneously identify at-risk and advanced beginning readers.

## *Objective of Dissertation Study*

The purpose of this dissertation was to use the one-parameter Rasch model for dichotomously-scored items to examine how measurement precision and time-efficiency could be optimized to enhance the classification accuracy of a screening measure designed to identify advanced and at-risk readers in second grade. This process started by analyzing the psychometric properties of a national reading test designed with classical test theory methods. Specifically, the responses from students in the normative sample (n = 801) that completed both forms of a subtest requiring word identification and vocabulary skills (n = 80 items) were assessed for dimensionality and then calibrated with the Rasch model using the Winsteps software program (Linacre, 2010). Person parameters were used to set cut points delineating the bottom 20% and top 20% of the second grade sample. The item parameters were used to develop five short-forms that varied by length (n = 10, 16, 22, 28, 34 items). Each short form was designed to maximize item information, and therefore person measurement, around the established cutpoints. The relative efficiency of each short-form was compared to both full-length subtest forms (n = 40 items each) with descriptive statistics and classification accuracy measures that include correlation coefficients, Cohen's kappa, sensitivity, and specificity. The goal of this process was to develop a universal screening measure that second grade teachers can use to enhance RtI frameworks by (1) more quickly and accurately identifying exceptional readers that will need differentiated instruction to be successful in compliance with federal mandates while simultaneously (2) establishing three instructional reading groups (advanced, on-track, at-risk) in order to drive the academic development and achievement of *all* students.

# CHAPTER II: LITERATURE REVIEW

In order to provide a framework for the present investigation, this chapter reviews five bodies of literature. These include (1) Identification and Intervention Models for At-Risk Readers, (2) Incorporating Advanced Readers into a RtI Framework, (3) Classical Test Theory, (4) Item Response Theory, and (5) Applying the Rasch Model.

## *Identification and Intervention Models for At-Risk Readers*

### Dual Discrepancy Model

Two models can be used to identify students with a learning disability (LD) in the United States. Most students considered for special education eligibility receive IQ and achievement tests. If there appears to be a significant "dual discrepancy" between IQ score and academic performance, the student is generally qualified as LD and admitted for special education services. The use of the IQ-discrepancy model for the identification of students with LD has come under widespread criticism and is increasingly referred to as the "wait to fail" model (Francis et al., 2005). Experts contend that requiring extensive assessments to diagnose LD as a prerequisite to intervention does not necessarily equate to better student outcomes and often results in long delays in determining eligibility and, therefore, providing services. Meanwhile, many of these measures have little instructional relevance (Fletcher, Coulter, Reschly, & Vaughn, 2004; Francis et al., 2005). A growing number of teachers and parents are similarly dissatisfied. A survey conducted by the National Center for Learning Disabilities (2002) found that 54% of parents and 72% of teachers felt that current identification methods for LD took too long to identify students in need and provide intervention. The Response to Intervention (RtI) identification model

was developed as an alternative to overcome many of the shortcomings inherent in the dual discrepancy model.

## Response to Intervention (RtI) Model

RtI is a multi-tiered educational service delivery model designed to improve student learning by allocating high quality instructional resources to students identified as at-risk for academic failure. There are two primary goals. The first goal is to provide early intervention services for students who are struggling in the general education curriculum before academic problems develop into acute deficits difficult to overcome. The second goal is to distinguish between students with reading problems due to a history of poor instruction and those with actual learning disabilities (Johnson, Jenkins, & Petcher, & Catts, 2008).

Most RtI models are comprised of three tiers, which include the general education classroom (Tier 1), an intermediary remedial class (Tier 2), and the special education classroom (Tier 3). To optimize learning progress, instruction at successive tiers is more intense and explicit while group size is reduced (Vaughn & Linan-Thompson, 2003). Since identifying children with reading deficits is the first step in preventing long-term academic failure, screening all students for beginning reading problems initiates the first tier of most RtI models. After initial screening is completed and potentially at-risk students are selected for Tier 2 intervention, movement between the tiers is driven by content mastery and growth rate on progress monitoring measures (Vaughn & Linan-Thompson, 2003). Progress monitoring involves the use of brief, rate-based measures delivered on a weekly, bi-weekly, or monthly basis to monitor mastery of specific instructional objectives and to inform instruction. As an illustration, students that demonstrate unsatisfactory progress in the general education classroom based on the results of a screening measure in Tier 1 receive more intensive and individualized

instruction in Tier 2. In most RtI models, Tier 2 involves small group tutoring sessions with systematic evidence-based instructional practices. Sessions typically occur at least four times per week and last between 20 and 40 minutes, depending on need. After 10 weeks, students are assessed. Those that did not master the established benchmarks receive another 10 weeks of Tier 2 intervention. Students that do not respond to Tier 2 intervention after 20 weeks become candidates for Tier 3 special education services, which is even more intensive and individualized. Students are identified with LD when their response to effective instruction is dramatically inferior to that of peers (Vaughn & Fuchs, 2003). Through this process it is possible to differentiate between students with a true learning disability and those that are under-achieving due to poor instructional practices.

## *Incorporating Advanced Readers into a RtI Framework*

All students with unique learning needs could benefit from instruction matched to instructional needs. As stated by Kurns and Tilly (2008), "If the needs of all students are going to be addressed, advanced or gifted students also need to be included in the data analysis" (p.21). It therefore stands to reason that it is just as important to identify academically advanced students as it is to identify those that are behind since both of these populations will need differentiated instruction in order to develop their academic skills. Just as second grade students that have not yet learned the alphabet should not be subjected to oral story reading exercises, second grade students that can read independently at the fourth grade level should not be subjected to reading instruction focused on correctly pronouncing individual letter sounds. Until we respect and address

the instructional needs of every student, it will be impossible to truly leave no child behind in the public education system.

Despite substantially rising inflation-adjusted per-student spending on K-12 education, Walberg (2001) found that U.S. schools ranked last in four of five international comparisons of educational progress in reading, science, and math through eighth grade. In the fifth case, they ranked second to last. For many years, the poor performance of U.S. students on international educational comparisons was explained away as misleading and a consequence of nationwide diversity. However, a recent analysis that disaggregated results by State prior to international comparisons showed that most States were performing at the bottom of the scale compared to other developed nations and none were performing in the very top. When averaged, U.S. students ranked 31st out of 56 countries in the percentage of students performing at a high level of accomplishment (Ripley, 2010). This should come as little surprise since the education of highly capable students is not prioritized by the federal government – as indicated by the absence of federal mandates protecting their rights to a fair and appropriate education similar to the rights enjoyed by students with disabilities. Yet, one could argue that being highly capable in American public schools is a disability because students that are performing several grade levels above expectations often suffer through frustration and boredom while experiencing retarded intellectual growth similar to their peers performing several grade levels below expectations. As noted by Sanders (1999),

> It is not our lowest-achieving children whom our system serves worst. It's our early high achievers among minorities. In Tennessee, for instance, the children who are getting hammered the hardest are the early high-achieving African American children. They do well in the early grades but decline in later grades. This comes from their higher likelihood of being in a succession of classrooms where the instruction is geared to lower achievers. Any children who have a likelihood of being in such an environment will experience what I call a shed pattern: declining like the roof of a country shed.

One major policy issue that currently confronts gifted education is deciding how eligibility should be determined. Similar to special education, many argue that the identification process is biased against minority and low socio-economic status students. According to Ford (1998), the majority of explanations for under-representation can be categorized as (1) recruitment issues related to screening and identification procedures, (2) personnel issues (e.g., teacher training and expectations); or (3) retention issues (e.g., student-teacher relations, learning environment). For the purposes of the current investigation, the remaining discussion will be limited to recruitment issues related to screening and identification procedures.

With an emphasis on universal screening, matching instructional services to instructional needs, and consistently monitoring academic progress, RtI could readily be used to advance the education of highly capable students and overcome many of the current challenges related to gifted education (Coleman & Hughes, 2009). Some further argue that by linking education policies for highly capable students to RtI and other special or general educational practices, the field can reach consensus on policy issues that could serve as a template for overall instructional and student improvement (Brown & Abernathy, 2009).

## *Universal Screening Methods and Evaluation*

Universal screening measures are designed to discriminate between individuals that do or do not have a particular condition. In education, the goal of using these instruments is to predict a negative outcome months or years in advance of the outcome so that teachers can intervene early and hopefully prevent the negative outcome. Risk decisions are made by selecting critical cut-points along a continuum of scores. The criterion for cut

scores delineating between poor, basic, and advanced performance is typically defined by a specific percentile (e.g., below 20th percentile) that corresponds to a test score (Johnson, Jenkins, Petscher, & Catts, 2009).

Universal screening measures for beginning readers consist of brief assessments with items targeting discrete skills that are highly predictive of later reading outcomes. Beginning reading constructs with the strongest predictive validity include phonemic awareness, letter knowledge, word identification, and reading fluency (O'Connor & Jenkins, 1999). Research further suggests that assessments of expressive and receptive vocabulary, sentence imitation, story recall, working memory, and attention may also have predictive value in forcasting reading problems (Catts, Fey, Zhang, & Tomblin, 2001; McCardle, Scarborough, & Catts, 2001). Consequently, most universal screening measures in beginning reading target the assessment of these skills.

Jenkins (2003) notes that beginning reading screeners should satisfy three criteria. First, the screen should accurately classify students as *at-risk* or *not at-risk* for reading failure. Second, the screen should be easy to administer, score, and interpret. And finally, the screen should demonstrate consequential validity such that the overall net effect for students is positive (Messick, 1989). According to Jenkins, this means that students identified as at-risk for failure must receive timely and effective intervention without any other students or groups being adversely impacted. However, beyond identifying students that will need remedial instruction to be successful, the most useful screens should also be able to identify students that will need advanced instruction to be successful. Though popular screening measures in beginning reading do not currently target the identification of advanced learners, teachers would clearly benefit from such information. Therefore, the screening measure developed in the current investigation was designed to simultaneously identify at-risk and advanced readers.

Binary classification analysis is often used to evaluate the utility of academic screening instruments. As noted in Chapter 1, this analysis involves the calculation of the test's *sensitivity* and *specificity*. In the current context, the sensitivity statistic describes how accurately the test identifies children with reading problems while the specificity statistic describes how accurately the test identifies children without reading problems. When an assessment is used to predict a binary outcome (e.g., deficient or proficient), four results are possible: true positive, true negative, false positive, or false negative. The term *positive* always indicates the presence of a problem while the term *negative* always indicates the absence of a problem. In the current context, (1) *true positives* are represented by students identified as having a reading problem according to the screening measure and the criterion measure; (2) *false positives* are represented by students identified as not having a reading problem according to the screening measure, but there is a problem - according to the criterion measure; (3) *true negatives* are represented by students identified as not having a reading problem according to the screening measure and the criterion measure; and (4) *false negative*s are represented by students identified as not having a reading problem according to the screening measure, but there is a problem - according to the criterion measure. Given these classifications, the sensitivity and specificity indices can be determined. The sensitivity index is calculated by dividing the number of true positives by the sum of true positives and false negatives. The specificity index is calculated by dividing the number of true negatives by the sum of true negatives and false positives (Gredler, 1997).

In theory, a perfect screen would differentiate between every child that does or does not have a reading problem with 100% accuracy. In practice, there seems to be a trade-off between sensitivity and specificity - as one increases the other often decreases. More specifically, raising the cut-point typically increases sensitivity, but decreases specificity.

In this case, the screening measure over identifies students as at-risk who are not really at-risk. Conversely, lowering the cut-point typically increases specificity, but decreases sensitivity. In this case, the screening under identifies students with true problems. For instance, perfect specificity could be achieved by setting the cut score of an academic screener to 0. Since all students would score 0 or above, the measure would correctly identify all students with satisfactory reading skills, however those with poor reading skills would not be correctly identified (i.e., poor sensitivity). As noted in the first chapter, the acceptable levels of sensitivity and specificity vary by field and according to the intended purpose of the test. In the field of special education, recommendations vary from at least 0.75 for both indexes (Gredler, 2000; Kingslake, 1983) to at least 0.80 for both indexes (Carran & Scott, 1992). If the purpose is to ensure that truly at-risk students are identified in the area of beginning reading, some suggest a more rigorous sensitivity standard of 0.90 with a specificity level close to 0.90 (Jenkins, 2003).

## *Review of Popular Screening Measures in Reading*

Most widely-used screening measures for early reading were influenced by the findings of the National Reading Panel (National Institute of Child and Human Development, 2000) and related empirical research. Several studies have examined the sensitivity and specificity of widely-used screening measures for second grade readers. The findings of these investigations are summarized below.

The classification accuracy of the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 2002) was analyzed in two studies. The DIBELS was developed using classical test theory methods. It is arguably the most popular screening measure in the United States. To date, the battery has been adopted by more than 14,000

elementary schools (https://dibels.uoregon.edu/data/index.php) and used to assess more than 1.8 million students (Samuels, 2007). The DIBELS consists of six standardized, individually-administered subtests: Letter Name Fluency, Initial Sound Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Oral Reading Fluency, and Retell Fluency. Based on research showing that fluency is a better predictor of reading success than accuracy alone, performance is measured in one-minute increments and indexed as rate per minute (Good, Simmons, & Kame'enui, 2001). Since the development of the DIBELS was financed by a federal grant, the measures are available to teachers via the Internet at no cost.

Riedel (2007) examined how well three *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 2002) subtest scores collected at the end of first grade (n = 1,518 students) predicted reading ability at the end of second grade. The Reading Comprehension subtest of the *Terra-Nova* (CTB/McGraw-Hill, 2003) was used as the external criterion measure. The Terra-Nova is a standardized, group-administered test with questions presented in multiple-choice format. The outcomes revealed that the sensitivity and specificity levels from the Phoneme Segmentation Fluency (59% and 59%, respectively), Nonsense Word Fluency (67% and 64%, respectively), and Oral Reading Fluency (71% and 71%, respectively) subtests were not sufficiently accurate predictors of reading success.

Johnson, Jenkins, Petscher, and Catts (2009) reached similar conclusions two years later. They examined how well the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 2002) and the *Peabody Picture Vocabulary Test* (PPVT; Dunn & Dunn, 1981) predicted outcomes on the S*tanford Achievement Test-10th Edition* (SAT-10; Harcourt Assessment, Inc., 2002). The PPVT is a measure of receptive vocabulary. Students are prompted to select one of four pictures that best depicts the

meaning of a word spoken by the examiner. Stimulus items are ordered from easy to difficult. Data were collected from a representative sample of 12,055 students beginning in kindergarten and continued through the end of their third grade year. In the end, the authors reported "the most significant limitation with the screeners (DIBELS subtests and PPVT) in this study is their lack of precision to identify most of the truly at-risk students without over-identifying very large numbers of students who are not at-risk" (p.184).

Foorman et al. (1998) evaluated the classification accuracy of students using a screening task included in the *Texas Primary Reading Inventory* (TPRI; Texas Education Agency, 2004). The TPRI was developed using the two-parameter item response theory model for dichotomously-scored data. There are two primary components to the test: a diagnostic assessment and a universal screening measure. The screening measure includes the Graphophonemic Knowledge subtest and the Phonemic Awareness subtest. It was designed to over-identify students with the rationale that false negative errors are more serious than false positive errors when identifying children with reading problems (Foorman & Ciancio, 2005). In the study conducted by Foorman et al. (1998), the Broad Reading cluster of the *Woodcock-Johnson Test of Achievement - Revised* (WJ-R; Woodcock & Johnson, 1989) was used as the criterion measure for second grade students. The cluster includes measures of word recognition skills and a clozed-based reading comprehension task. Risk of reading failure for the 537 students tested in the beginning of second grade was defined as a criterion of 0.5 grade equivalents below grade level (<36th percentile) on the WJ-R Broad Reading at the end of second grade. The results revealed impressive levels of sensitivity (91%) and specificity (85%).

The screening instruments analyzed in these studies were designed to measure aspects of phonological awareness, print knowledge, or oral vocabulary. Across studies, the screening assessments that targeted letter-word identification skills were consistently

the most precise followed by assessments that targeted decoding skills (i.e., nonsense words). Meanwhile, the instruments that targeted phonological awareness skills were consistently the least accurate across studies. The outcomes further suggest that screens composed of more than one measure are more accurate than screens that measure only one skill (Johnson, Jenkins, & Hudson, 2007).

## Improving the Classification Accuracy of Screening Measures

Two methods can be used to improve the classification accuracy of a screening measure. The first method is to adjust the cut-point for the measure and recalculate the binary classification indices. The second method is to improve the structure of the scale using item response theory models (Schultz-Larsen, Kreiner, & Lomholt, 2007).

Test theories are important to the study of psychological dimensions because they provide a framework for understanding behavior and exploring the effects of education. Two distinct theories of psychological measurement can be used to develop academic assessments and screeners: classical test theory or item response theory. While valid and reliable measures can be developed within both frameworks, there are unique benefits of using item response theory models in educational contexts which include: (1) item parameters are not dependent on the group(s) from which they were originally estimated; (2) scores estimating student ability are not related to test difficulty; (3) shorter tests can be created that are more reliable than longer tests; and (4) item statistics and ability estimates are reported on the same scale (Embretson & Reise, 2000). In an effort to establish the context for the methods used in the current investigation, the basic principles, fundamental assumptions, test development techniques, and primary limitations of classical test theory and item response theory models will be reviewed below.

## *Classical Test Theory*

### General Principles

Classical test theory has been the predominant measuring system used to make inferences about latent ability based on observable behaviors since the turn of the century. Gulliksen (1950) was the first to synthesize the principles and technical advancements in classical test theory into a single comprehensive source. Accordingly, his work will be the primary reference in the following discussion.

Classical test theory rests on aspects of a total test score comprised of multiple items. The theory assumes that the raw test score ($X_i$) obtained by an individual is comprised of a true component ($T_i$) and a random error component ($E_i$). This fundamental premise is denoted

$$X_i = T_i + E_i \quad , \qquad (1)$$

where $X_i$ is the observed raw test score, $T_i$ is the true score, and $E_i$ is the error score for person *i* on the test. The observed score is defined as the raw score an individual receives on the test and serves as the basis of individual ability estimates. The true score ($T_i$) is a hypothetical construct derived by taking the mean score that the individual would get on the same test given an infinite number of testing sessions. As random error ($E_i$) decreases, the extent to which the observed score ($X_i$) reflects the true score ($T_i$) increases. Therefore, the overarching goal of test developers using classical methods is to measure true skill proficiency while minimizing and effectively coping with random error (Kline, 2005).

#### FUNDAMENTAL ASSUMPTIONS

Three fundamental assumptions are required to develop an assessment using classical test theory. Each assumption relates to the treatment of error scores. First, error scores are assumed to be random in nature and thus are defined as normally distributed with a mean of zero. Second, error scores are uncorrelated with each other. In other words,

there is no systematic pattern to score fluctuations over time. Third, it is assumed that error scores are uncorrelated with the true scores on any given test, true scores on any outside criterion measure, and error scores on parallel forms. Parallel forms are defined as tests that measure the same latent skill proficiency, yield the same true score for a given individual, and have the same conditional standard error of measurement. Given the definitions of error scores and parallel forms, a person's true score can be theoretically derived by taking the difference between the observed score and the error score with the equation

$$T_i = X_i - E_i \quad . \quad (2)$$

The variance of the observed score can be expressed as the sum of the variance of the true score and the variance of the random error scores denoted

$$s_x^2 = s_t^2 + s_e^2 \quad . \quad (3)$$

The reliability coefficient is defined as the correlation coefficient between observed scores on two parallel forms of the test. Through the following formula it can be shown that the reliability coefficient of the test equals the ratio of the true score variance to the observed score variance

$$r_{xx} = s_t^2 \div s_x^2 \quad . \quad (4)$$

Thus, the variance of the true scores can be obtained from observable test scores by rewriting Equation 4 as follows

$$s_t^2 = s_x^2 r_{xx} \quad . \quad (5)$$

Substituting the observable equality for the true score variance in Equation 3 results in the following equation

$$s_x^2 = s_x^2 r_{xx} + s_e^2 \quad . \quad (6)$$

Given all the terms except the one of error score variance can be obtained from the observed test scores, the variance of the error scores can be obtained by rewriting

Equation 6 to solve for the variance of the error scores. The resulting equation is

$$s^2_e = s^2_x - s^2_x r_{xx} \quad . \qquad (7)$$

Simplifying Equation 8 and taking the square root of both sides of the equation yields the following formula for the standard deviation of the error scores

$$s_e = s_x\sqrt{1 - r_{xx}} \quad . \qquad (8)$$

The standard deviation of the error scores, also referred to as the standard error of measurement, is defined as the distribution of random errors around the true scores. It indicates how accurately a trait is assessed by a measure. As shown in Equation 8, the standard error of measurement is a function of the variability of the observed test scores and the reliability coefficient of the test for a given population. As such, it can be used to describe the distribution of the observed scores for one individual or groups of individuals with a given true score. In both contexts, it is expected that some observed scores will be higher than the true score while others will be lower than the true score due to random measurement errors. More specifically, it is expected that approximately 68% of the observed scores will fall within +/-1 standard error of measurement from the true score while approximately 95% of the observed scores will fall within +/-2 standard error of measurements from the true score. In this way, the standard error can be used to make inferences about an individual's true score based on knowledge of the observed test score. In practice, the statistic is expressed as a band of errors surrounding the observed score in which the true score is expected to reside with an approximate degree of confidence. This confidence can be increased to approximately 95% by multiplying the *z* value associated with the 95% confidence interval ($z = 1.96$) by the value of the standard error. This value is then added to and subtracted from the observed score to derive the confidence interval which is expected to contain the true score (Dodd, 1985).

## ITEM ANALYSIS TECHNIQUES

In classical test theory, the raw test score is used to determine how well the student performed while the *p* value (probability of correct response) and item-total correlation coefficient are used to determine how well the test items performed. Standard item analysis practices involve the examination of item means, variances, standard deviations, difficulty indices, and discrimination indices. The mean of a dichotomous item is equal to the proportion of individuals who answered the item correctly (denoted *p*). The variance of each item can be calculated by multiplying $p \times q$, where *q* is the proportion of individuals who did not answer the item correctly. Thus, the standard deviation of a dichotomous item is the square root of $p \times q$. For instance, if 100 examinees answer an item correctly and 500 answer incorrectly, then the *p* value for the item is 100/500, or 0.2. The *q* value is 0.8 (1.0 - 0.2 = 0.8). The variance of the item is 0.16 (0.8 x 0.2 = 0.16). And the standard deviation is the square root of 0.16, or 0.4 (Kline, 2005).

Beyond a descriptive statistic, the *p* value is used as the index of item difficulty. High values denote easy items while lower values denote more difficult items. Items with *p* values of 0.5 (i.e., 50% of the group passed the item) provide the highest levels of differentiation between individuals in the group. Thus, *p* values closest to 0.5 discriminate best among examinees (Kline, 2005).

After item difficulty is established for a set of test items, item-total correlations are typically calculated to measure how well individual items discriminate between high-scoring and low-scoring examinees. With dichotomously-scored items, the statistic is most often expressed as the point-biserial correlation coefficient between individual items responses and the total test score. The item-total correlation describes the discrimination power of the item.

### Item Selection Techniques

The equidiscriminitive item-total correlation technique is frequently used to select test items. With this technique, test developers select items with the highest correlation with the overall score by calculating the item-total correlations for three ability-based subgroups stratified by total score. The highest ranking items should be able to detect skill proficiency across the entire range of scores (Kline, 2005).

The optimal level of item difficulty depends on the anticipated ability distribution of the target population. Most norm-referenced tests are designed to differentiate between students across the range of ability levels. For this purpose, Anastasia and Urbina (1997) recommend that the average item difficulty should approximate 50% with a fairly large dispersion while noting that items distributed between 15% and 85% are typically considered acceptable. Given a set of items with acceptable discrimination and difficulty values, items are selected to satisfy the conditions of the table of specification. An attempt is always made to choose the items with the highest discrimination parameters. Items with a discriminating power of at least 0.2 are generally considered appropriate for longer test, however more discriminating items should be used with shorter tests (Bernstein, 1994).

### Primary Limitations

Despite the popularity of classical test theory, there are several major conceptual limitation. Perhaps the most significant is that item parameters and ability estimates are interdependent. As discussed in Chapter 1, the difficulty and discrimination power of test items are a function of the test scores of the sample population; yet the test scores of the sample population are a function of the difficulty and discrimination power of the test items. More specifically, tests administered to samples of examinees with above-average ability will result in higher item difficulty values while tests administered to samples of examinees with below-average ability will result in lower item difficulty values.

Meanwhile, tests administered to heterogeneous groups will result in higher discrimination values while lower values will be obtained from homogeneous groups. Even with elaborate sampling methods to ensure a representative sample during the norming process, student characteristics and test characteristics cannot be separated and ability estimates can only be safely interpreted relative to the particular population of examinees with which the item indices were originally obtained (Dodd, 1985). The magnitude of the interdependency can be examined with the standard error of measurement conditioned on ability. For instance, the WISC-R full scale IQ test has as standard error of measurement of 2.96 for 12 1/2 year-old children and a standard error of measurement of 3.23 for 13 1/2 year-old children, however the reliability coefficient for both populations is 0.96 (Wechsler, 1972). The difference in the standard error of measurement across both populations is necessarily a function of the variability in observed scores within each population. Therefore, even IQ scores derived from tests built with classical methods are not standard and absolute units of measurement. Rather, they must be defined relative to the variation of observations across groups (Dodd, 1985).

The second limitation is related to the general treatment of the standard error of measurement. In classical test theory, it is assumed to be constant for all ability levels across all subgroups of the population (Kline, 2005). Yet, in practice, research shows that measurement error is generally higher at the extremes of the ability distribution (Kline, 2005) and can also vary across sub-populations (Wechsler, 1972).

The third limitation of classical test theory methods is that ability estimates are calculated as the sum of the correct individual item scores. As such, raw test scores are on a discrete ordinal scale, yet equal distance between observed scores may not denote equal differences in cognitive functioning across the score range. Consequently, attempts to

interpret the clinical meaning of a score or score improvement are necessarily confounded (Kline, 2005).

The final limitation of classical test theory methods is that assessments cannot be readily adapted as screening measures or designed to optimize measurement precision for specific populations (e.g., low-ability or high-ability students). Test developers generally rely on Cronbach's alpha, item-total correlations, and expert opinions to shorten assessments (see Coste et al. for a review of methods). The problem of such approaches is that the scores on the abbreviated forms are not directly comparable with the scores on the unabbreviated form because they are not on the same scale.

## *Item Response Theory*

### General Principles

Item response theory (IRT) was developed to overcome the vexing interdependency issues inherent to classical test theory methods. The primary benefits of using IRT models include: (1) item difficulty is not dependent on the group(s) from which they were originally estimated; (2) scores estimating student ability are not related to test difficulty; (3) shorter tests can be created that are more reliable than longer tests; and (4) item statistics and ability estimates are reported on the same scale (Embretson & Reise, 2000).

#### FUNDAMENTAL ASSUMPTIONS

Three fundamental assumptions are required to use most IRT models. First, it is assumed that a single ability accounts for differences in person responses to items (Embretson & Reise, 2000). Lord (1968) points out that while few tests are composed of items that are strictly unidimensional, many tests provide an adequate approximation of this assumption. Factor analytic techniques can be used to examine the dimensionality of

test items. Reckase (1979) states that the first factor should account for at least 20% of the total variance for the item parameters to be stable. Tests that do not meet this assumption can be subdivided into homogeneous subsets of items and then each subset can be analyzed separately.

The second assumption of IRT models is that test items have local independence. In theoretical terms, this means the probability of responding to each item is statistically independent of the probability of responding to any other item for examinees with the same ability. In practical terms, this means that the content provided in one item does not aide an examinee in answering another item. Given local independence, examinee ability can be estimated from any subset of items with ability estimates that are on the same original scale.

The final assumption is that a mathematical function can be derived to model the probability of a response to any given item conditional on ability level (Hambleton & Swaminathan, 1985). The mathematical function contains item characteristics (parameters) that allow the probability of a correct response for each ability level to be modeled graphically.

### Primary Limitations

Despite the many advantages of using IRT models, there are several limitations. First, large samples are needed. The number of participants needed depends on the model employed. Given the property of parameter invariance, it is not necessary to orchestrate a strict representative sample of the target population to calibrate test items. However, a large, heterogeneous sample is needed to ensure accurate estimation (Dodd, 1985). To obtain stable parameter estimates for dichotomously-scored items, the requisite sample sizes generally range from 300 to 3000 examinees depending on the particular item response theory model.

The second limitation is that the assumption required to use IRT models described above are much stricter than those required by classical test theory. Although IRT models are generally robust to modest violations (Drasgow & Hulin, 1990), explicit tests of the assumptions must be conducted in order to analyze and interpret data with confidence.

The final limitation of using IRT models is that conducting the analyses and communicating the results to non-technically oriented audiences is a complex, time-intensive, and difficult process. Beyond a solid understanding of IRT models and principles, these analyses require specialized software. To date, the software is notoriously non-user friendly and relatively expensive.

## *Applying the Rasch Model*

### General Principles

Of all item response theory (IRT) models proposed for person measurement, the Rasch model (often referred to as the one-parameter logistic model) is considered to be the simplest and most efficient (Embretson & Reise, 2000). Like other IRT models, the Rasch model transforms ordinal data to an interval scale of scores for item difficulties and person abilities. The Rasch model is therefore probabilistic such that the distance between an item's difficulty and a person's ability governs the probability of being successful on any given test item. The basic premise is that there is a higher probability of correctly answering easier items than correctly answering more difficult items. In turn, given harder questions, there is a higher probability that more proficient individuals will answer correctly and a lower probability that less proficient individuals will answer correctly (Rasch, 1980).

### Dichotomous Model

Of all variations of the Rasch model, the dichotomous model is the simplest. It predicts the conditional probability of a binary outcome (correct = 1, incorrect = 0), given the ability ($\theta$) of the person ($n$) and the difficulty ($b$) of the item ($i$). This requires the estimation of one ability parameter for each person ($\theta_n$) and one difficulty parameter for each item ($b_i$). Similar to classical test theory techniques, person ability is initially calculated as the proportion of items on which each person succeeded and item difficulty is initially calculated as the proportion of the sample that succeeded on each item. Rasch analysis software programs can be used to transform the raw percentage scores from an ordinal scale to a *log*arithmic scale based on the *odds* of success. The resulting item difficulty ($b_i$) and person ability ($\theta_n$) estimates are expressed in logits on a scale of log odd ratios, which is referred to as a logit scale (Bond & Fox, 2007).

### Fundamental Assumptions

In addition to the three assumptions common to most IRT models, there are additional assumptions unique to the Rasch model. First, it is assumed that all items are equally discriminating. In other words, some items do not differentiate better or worse for examinees at a given ability level. Second, the set of test items need to measure a single, or unidimensional, construct such as word identification skills. If these assumptions hold, the Rasch model will order items and persons on an interval scale and the raw score can be used to estimate person ability (Rasch 1980).

Similar to classical test theory, there is a one-to-one correspondence between a person's test score and his or her ability estimate with the Rasch model. This means that, given a set of test items, people with the same raw score will have the same ability estimate in logits regardless of which items were answered correctly. The rationale is that if all items have equal discrimination power, then each item should have the same weight

in determining ability (Bond & Fox, 2007).

### PARAMETER ESTIMATION

While ability parameters are initially estimated as the proportion of items on which each person succeeded, they are ultimately based on the response pattern of the person and the characteristics of the items (Bond & Fox, 2007). The challenge of estimation is to determine the difficulty parameter for each item and the ability parameter for each person when both ability and item parameters are initially unknown, but examinee responses are known. This is similar to regression analysis whereby parameters of the model (i.e., regression coefficients) must be estimated from observed responses to a variable. However, the independent variable can be observed in regression analysis while the regressor variable ($\theta$) in the Rasch model is unobservable (Bond & Fox, 2007).

The process of parameter estimation begins by constraining person estimates, then calculating the initial pass-versus-fail proportions for each item, and then using the item estimates to calculate the initial the pass-versus-fail proportions for each person. Through an iterative process known as the Newton-Rhapson procedure, the first round of ability estimates are conditioned upon to obtain modified item parameter estimates, which are then conditioned upon to obtain modified ability estimates. The iterations continue until either the calibrated values for both ability and item parameter estimates "converge" (i.e., stabilize) or until a pre-specified number of iteration cycles are completed. The resulting log-odds scale is an interval scale in which the unit intervals (logits) between person and item parameter locations have a consistent value (Bond & Fox, 2007).

At each ability level, there is a certain probability that an examinee with that ability will correctly answer the item. The probability ranges between 0 and 1. By comparing the person's estimated ability with the item's difficulty, the Rasch model will also predict the probability of that person answering that item correctly. For difficult test items, examinees

with higher ability will be more likely to answer correctly than examinees with lower ability. Given estimates of person ability ($\theta_n$) and item difficulty ($b_i$), the probability of correctly answering an item can be calculated with the formula

$$P_{ni}(x=1) = f(\theta_n - b_i) \quad , \tag{9}$$

where $P_n$ is the probability, $x$ is a given item score, and 1 is a correct response. The equation states that the probability ($P_n$) of person $n$ getting a score ($x$) of 1 on item $i$ is a function ($f$) of the difference between the person's ability ($\theta_n$) and the item's difficulty ($b_i$).

Given $\theta_n$ and $b_i$, Equation 1 can be expanded to show that the function ($f$) expressing the probability of a correct response consists of a natural logarithmic transformation of the person ($\theta_n$) and item ($b_i$) estimates. This relationship is mathematically denoted

$$P_{ni}(x_{ni} = \frac{1}{\theta_n, b_i}) = \frac{e^{(\theta_n - b_i)}}{1 + e^{(\theta_n - b_i)}} \quad , \tag{10}$$

where $P_{ni}(x_{ni} = 1/\theta_n, b_i)$ is the probability of a correct response ($x = 1$) by person $n$ on item $i$, given person ability ($\theta_n$) and item difficulty ($b_i$). This probability is equal to the constant e, or natural log function (2.7183) raised to the difference between the person's ability and the item's difficulty, and then divided by 1 plus this same value. For example, if a person's ability estimated at 2 logits ($\theta = 3$) was given an item with a difficulty of 2 logits ($b = 2$),

the expected probability of that person correctly answering that item is denoted

$$P_{ni}(x = \frac{1}{\theta(3), b(2)}) = \frac{2.7183^{(3-2)}}{1 + 2.7183^{(3-2)}} = \frac{2.7183^{(1)}}{1 + 2.7183^{(1)}} = 0.73 \quad (11)$$

In this case, the person has a 73% chance of answering the item correctly. If that same person ($\theta$ = 3) was given an item with a difficulty estimate equal to ability ($b$ = 3), the expected probability of correctly answering the item is mathematically derived

$$P_{ni}(x = \frac{1}{\theta(3), b(3)}) = \frac{2.7183^{(3-3)}}{1 + 2.7183^{(3-3)}} = \frac{2.7183^{(0)}}{1 + 2.7183^{(0)}} = 0.50 \quad (12)$$

This pattern shows that as items become progressively more difficult, the probability of correctly answering the test item becomes increasingly more difficult. And when the person's ability estimate equals the item's difficulty location on the scale, then the person has a 50% chance of correctly answering the test item. It also illustrates that, similar to classical test theory, the raw score is a sufficient statistic to estimate person and item parameters.

Table 1 shows the probabilities of correctly answering a dichotomously-scored item across a range of difference values between $\theta_n$ and $b_i$. Note that as person ability ($\theta_n$) increases, the probability of correctly answering the item ($b_i$) also increases.

TABLE 1: Probabilities of Correctly Answering a Dichotomously-Scored Item with the Rasch Model

| $\theta_n - b_i$ | $P_{ni}(x = 1)$ |
|---|---|
| -3.0 | 0.05 |
| -2.0 | 0.12 |
| -1.0 | 0.27 |
| 0.0 | 0.50 |
| 1.0 | 0.82 |
| 2.0 | 0.88 |
| 3.0 | 0.95 |

Plotting the probabilities of a correct response conditioned on ability yields the S-shaped curve depicted in Figure 1. The probability of a correct response is represented along the y-axis. And the range of item difficulty and person ability values is placed on the same scale along the x-axis. Notice how the probability of a correct response starts near zero at the lowest level of ability and approaches one until it reaches the highest level of ability. Accordingly, the S-shaped curve for each item, referred to as the *item characteristic curve*, is a mathematical function that describes the relationship between latent ability and the probability of a correct response.

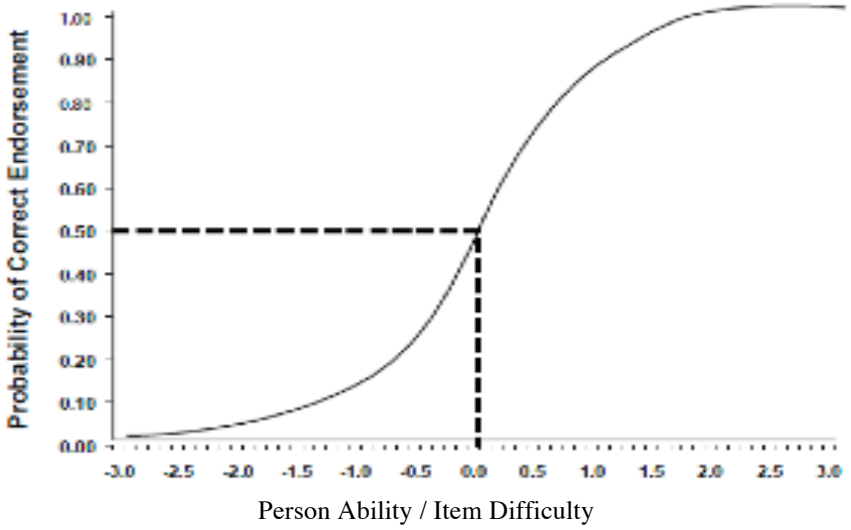


*FIGURE 1. Item Characteristic Curve for the Rasch Model*

The item characteristic curve is the foundation of all item response theory models. It can be viewed as the nonlinear regression of item scores relative to the underlying latent variable (θ), which is usually assumed to have a standardized normal distribution with a mean of 0 and a standard deviation of 1 (Lord, 1980). Increasingly negative values along the x-axis indicate easier items and lower ability examinees while increasingly positive values indicate harder items and higher ability examinees. A value of 0 reflects an average item difficulty and person ability level. For the item characteristic curve depicted in Figure 1, students with an ability of -3.0 would have an approximate 1% probability of correctly endorsing the item while those with an ability of +3.0 would have an approximate 99% probability of correctly endorsing the item.

One parameter drives the general form of the item characteristic curve: item difficulty. The item difficulty parameter is a location index that delineates how well the item measures examinees with different abilities. The Rasch model defines item difficulty

as the value on the x-axis at which the probability of correctly endorsing the item is 0.50. The difficulty of the item shown in Figure 2 is 0.0.

The more difficult an item is, the further the curve shifts to the right along the x-axis. Figure 2 depicts three items with equal slope, but different difficulty parameters.

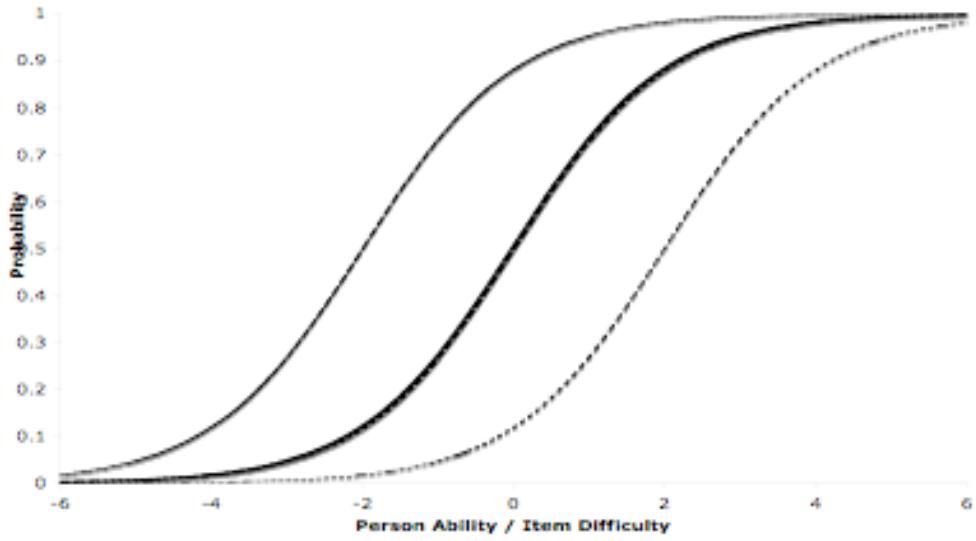


*FIGURE 2: Item Characteristic Curves for Three Dichotomous Items*

## Item Analysis Techniques

Fit statistics are used to detect unexpected response patterns that do not conform to the Rasch measurement model. Two chi-square statistics are commonly examined: the infit mean square (infit MNSQ) and the outfit (MNSQ). The infit MNSQ represents the information-weighted mean square residual difference between observed and expected responses. The infit statistics are sensitive to unexpected responses near a person's ability level. The outfit statistic is the usual unweighted mean square residual and is more sensitive to outliers. High infit and outfit statistics reflect underfit, which means lack of predictability. Low infit and outfit statistics reflect overfit, which means over-predictability. According to Bond and Fox (2007), outfit or infit mean-squares greater than or equal to 2.0 can distort measurement. Items with such extreme values should be re-examined and possibly deleted from the final item pool.

### Item Selection Techniques

Final item selection depends on the amount of information the item contributes to the overall measurement precision of the test. The most common item selection procedure is maximum information selection, where items that provide the most information near the targeted ability estimate is selected (Birnbaum, 1968; Lord, 1977). Three unique features of IRT models are used to make this determination: (1) item information functions, (2) test information functions, and (3) standard error of measurement conditional on ability level for each item.

Item information refers to the precision of measurement that an item provides for each ability level, with higher information denoting more precision (Embretson & Reise, 2000). Since test items do not measure all ability levels with equal precision, information

is not consistent across the scale. The item information function is denoted

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)(1 - P_i(\theta))}, \qquad (13)$$

where $P_i(\theta)$ is the probability of a correct response to item $i$ conditioned on ability ($\theta$), and $P'_i(\theta)$ is the first derivative with respect to ability ($\theta$) (Embretson & Reise, 2000). In simpler terms, the item information can be derived by dividing the squared slope of the item characteristic curve at ability level θ by the squared standard error of measurement at θ. Items "function" best at the θ level that corresponds to the item difficulty. For example, easier items peak at lower ability values along the x-axis, which means they function as measures better for low ability examinees.

Figure 3 shows the characteristic curve and information function for the same item. Line A shows where persons with an ability of 1.0 have a 50% chance of answering the item correctly. This probability of a correct response occurs when the ability of the person matches the difficulty of the item. Correspondingly, 1.0 logits is also the value of the item difficulty parameter.

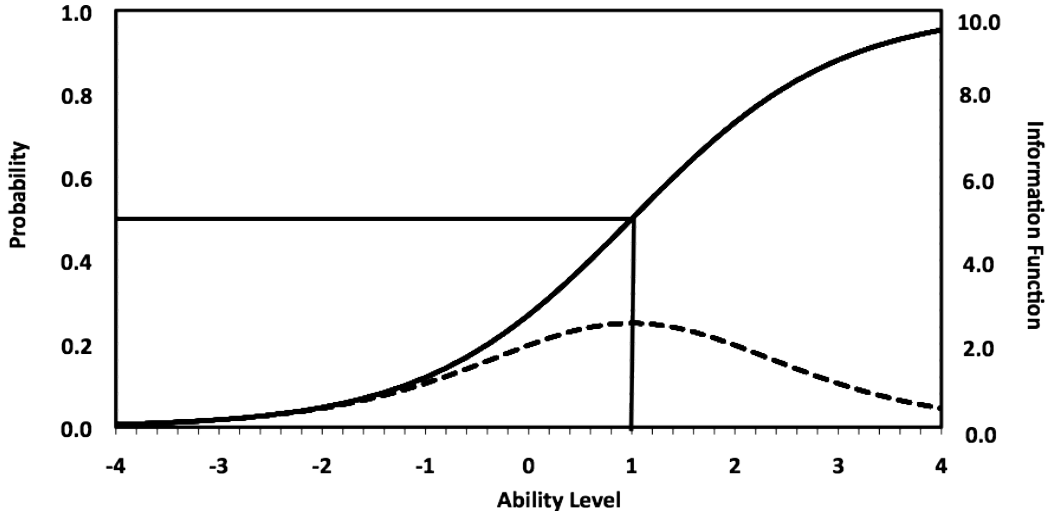


*FIGURE 3: Item Characteristic Curve and Item Information Function for the Same Dichotomous Item*

The test information function is the sum of item information functions. The additive property is possible given local independence among items (Lord, 1980). The test information function is denoted

$$TI(\theta) = \sum_{i=1}^{I} I_i(\theta) \quad . \qquad (14)$$

The test information function allows psychometricians to see the relative contribution of each item by determining the overall effectiveness of a set or sub-set of test items while estimating the effect of adding or deleting particular items (Embretson & Reise, 2000). Figure 4 depicts the TIF for a test with three content areas. The TIF is bimodal with peaks at the lower and upper end of the ability continuum.

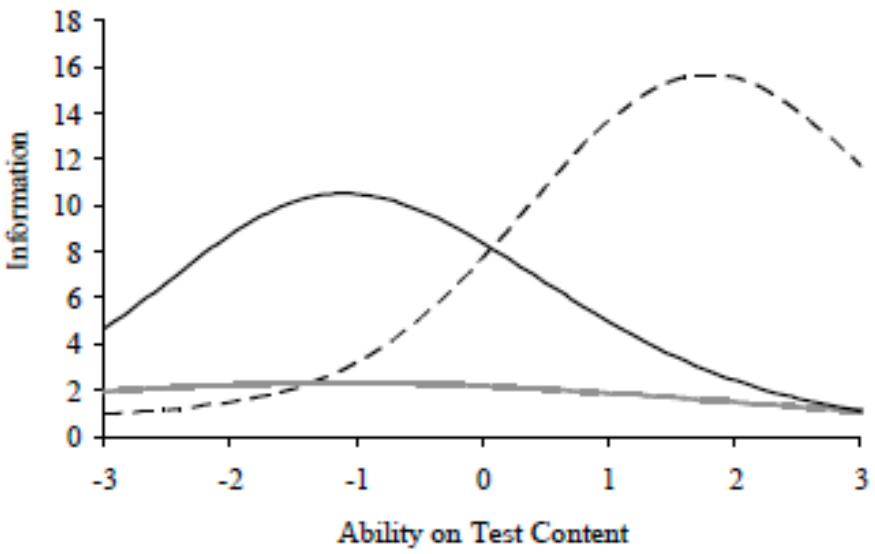


*FIGURE 4: Test Information Function for Scale with a Bimodal Distribution*

Measurement precision can be further examined with the standard error associated with a given ability level ($\theta$). It is calculated as the square root of the reciprocal of the test information function denoted

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad . \qquad (15)$$

To the extent the test items do not measure all ability levels with equal precision, the standard error is not consistent across the scale. The standard error tends to be highest at the extremes of the ability continuum and indicates where inadequate information is provided by the item parameters (Embretson & Reise, 2000).

In summary, item selection techniques rely on the simultaneous analysis of item information functions, test information functions, and the standard error conditioned on ability. Together, these features can be used (1) to determine the relative efficiency of

various test forms at different ability levels by comparing the informativeness of each using ratios of test information and error measurement curves and (2) to estimate how the addition or deletion of a single item will influence measurement precision (Embretson & Reise, 2000). Through this process, test items are selected to maximize information and minimize error at the ability levels targeted for measurement. The result is shorter forms of comprehensive assessments with reliability equal to or greater than levels achieved by classical test theory methods (Dodd, 1985).

## *Statement of Problem*

Screening measures in beginning reading play a pivotal role in identifying the students that will need differentiated instruction in order to be successful. As such, universal screening constitutes the foundation of the first tier of RtI frameworks. Problematically, research shows that many screening measures used to identify at-risk beginning readers are imprecise. As a result, too many students are being either misidentified or under-identified.

Item response theory models can be used to significantly improve assessments that target exceptional learners. The seminal investigation reported by Foorman and colleagues (1998) was the first to describe how beginning reading screening measures can be developed to identify at-risk students using item response theory. In their study, the maximum information method was used to select items for one short-form with the two-parameter logistic IRT model. To date, no studies have examined how measurement precision and efficiency can be optimized to increase the classification accuracy of a screening measure across forms that (1) vary by length using the Rasch model and (2) target at-risk and advanced learners.

This dissertation examined the classification accuracy of a succession of short-forms with a bimodal test information function using the one parameter Rasch model for dichotomously scored items. This study was designed to answer the following question:

1. What short-form length (n = 10, 16, 22, 28, 34 items) optimizes the time-efficiency and classification accuracy of a beginning reading screening measure as compared to both full-length forms (n = 40 items each) in terms of simultaneously identifying at-risk and advanced second grade readers?

# CHAPTER III: METHODOLOGY

The Rasch model was used to examine the psychometric properties of a diagnostic reading assessment developed with classical test theory methods. On the basis of the Rasch analysis, five short forms were created for use as screening measures. The maximum information item selection method was used to select items and optimize measurement precision around two cut points delineating the top and bottom 20% of second grade students in the normative sample. Measurement precision was examined and the classification accuracy of the item selection procedure across forms was analyzed.

## *Item Pool*

The present investigation utilized the normative dataset of a national reading assessment. The instrument was designed to identify students with reading problems in Kindergarten through Grade 6, diagnose specific skill deficits, and guide the instructional decision-making process. It was standardized on a nationally representative sample of 1,018 children aged 6 through 13 tested in the spring of 1999 and fall of 2002. Validity studies suggest the measure is correlated with other tests of reading, intelligence, language, and achievement.

The data used to obtain item parameters for the item bank consisted of examinee responses from both forms of a subtest requiring both word-identification and vocabulary skills. A total of 80 dichotomously-scored items ordered from easy to difficult were included across both test forms. There are two item formats which differ by visual stimulus (picture or word) and the number of response options. In the first format (n=16), each item consists of a picture and four words. Students are prompted to select the word that best represents the picture. In the second level (n = 64), each item consists of five

words (e.g., a, in, I, and, out). Students are prompted to select the two words that are opposites. The authors report high concurrent validity of the subtest scores with other established instruments such the WJ-R Word Attack (.88/.58) and Broad Reading (.78/.64) scores, however it should be noted that these findings are based on a study that included only 28 students.

## *Analysis of Dimensionality*

A principal components analysis was employed to assess the dimensionality of the item pool on the subset of the sample that completed both forms of the subtest. A principal components analysis is used to illuminate the structure of a relatively large set of variables and determine the number of factors that account for the total variance of the data. The a priori assumption is that any test item may be associated with any factor (latent variable) related to ability. A scree test is often generated to visually analyze data dimensionality. It is conducted by plotting the number of dimensions on the x-axis and the corresponding eigenvalues on the y-axis (Cattell, 1966). The objective of the plot is to identify the point at which the eigenvalues form a descending linear trend. Each data point located above the visual "elbow" in the scree test represents a factor in the data. Lord (1980) states that if the first eigenvalue is considerably greater than the second and the second eigenvalue has approximately the same magnitude as the other eigenvalues, then the items that comprise the scale can be considered unidimensional. Reckase (1979) further notes that the first factor should account for at least 20% of the common variance for item parameters to be stable.

## *Parameter Estimation*

Person and item parameters were calibrated with the Rasch model for dichotomously-scored data using the WINSTEPS software program (Linacre, 2010) and then examined for model fit. It is not possible to adequately estimate the ability of persons that answer all items correctly or incorrectly. All that can be concluded is that such individuals have either too little ability to score on the test or more ability than needed for the test. Similarly, it is not possible to adequately estimate the difficulty of items that all persons answer either correctly or incorrectly. Consequently, the results of any extreme persons and extreme items were omitted as inadequate before final parameter calibration (Bond & Fox, 2007).

## *Analysis of Model Fit*

For a measurement instrument to be useful, it's quantitative status must remain constant across contexts. The more applicable the model in a wide range of contexts, and the more useable the results, the more likely it will be to meet practical needs and form the basis of scientific progress (Linacre, 2003). Measures which are "generally objective" such that any two observers given the same observation will report back the same number as a measure - represents the gold standard which distinguishes a true scientific measurement instrument (Stenner & Horabin, 1992). In the case of the ruler, this means that the length calibrations will be maintained regardless of the object being measured (Stone, 2001). In the case of an academic test, this means that the test will be invariant such that it will (1) maintain the same level of difficulty regardless of who is taking it and (2) measure examinees with the same precision regardless of the difficulty of items included on it.

All Rasch analyses must include an evaluation of how well the data fit the measurement model in order to validate test items. Fit analysis evaluates how well the data cooperate with the construction of the measurement scale. The Rasch model requires that (1) more able examinees have a higher probability of correctly answering any given item than less able examinees and (2) all examinees have a higher probability of correctly answering an easier item than a harder item (Wright & Stone, 1999). Response plausibility, or fit, is calculated as the difference (B*n* - D*i*) between the estimates of person ability B*n* and item difficulty D*i* for each person *n* and item *i*. When this difference is positive it means that the item should be easy for the person. The more positive the difference, the easier the item is expected to be and hence the greater the probability that the person will succeed on that item. When the difference is negative, the item should be difficult for the person. The more negative the difference becomes, the more difficult the item should be for the person and hence the greater our expectation that the person will fail on that item (Wright & Stone, 1999).

The invariance principle requires that the relative difficulties of the items should remain stable across substantially different subsamples (Fox & Bond, 2007). Fit statistics were reviewed to detect departures from the measurement model. According to Bond and Fox (2007), outfit or infit mean-squares greater than or equal to 2.0 can distort measurement. Items and persons that failed to meet these criteria were further examined.

## *Short-Form Development*

The ability estimates for the second grade students without missing data included in the normative sample (n = 78) were ranked across the ability continuum. Two cut points were set to the θ values delineating the top 20% and bottom 20% of the sample given

ability estimates based on responses to the test items. This resulted in three performance categories: At-Risk, On-Track, and Advanced. Next, item information functions were used to select items for short-forms to create an optimal test information function. Items that provided the most information closest to the θ value delineating the at-risk group and the advanced group were selected. This method was repeated iteratively until the predetermined number of items for each form (n = 10, 16, 22, 28, 34 items) was satisfied. After each short-form was created, test information functions were calculated and analyzed.

Ability estimates were calculated for each short form with the WINSTEPS software program. The ability estimates were based on the real responses to the items included on each form. Using the raw score from both long-forms (i.e., Form A and Form B) as the reference, the discriminitive validity of the succession of subscales was assessed by calculating the (1) correlation with long-form scores, (2) classification consistency using Cohen's kappa (Cohen, 1960), and the degree to which each short-form detected at-risk and advanced 2nd grade readers with sensitivity and specificity (Altman & Bland, 1994).

The consistency of ability estimates across forms was assessed with the Pearson correlation coefficient (*r*). Since the short-forms measure the same aspect of reading as the long-forms, it was expected that they would be highly correlated. To investigate this hypothesis, the intercorrelations among the short-forms were calculated.

Cohen's kappa (*K*) was calculated to examine the effectiveness of each short-form in predicting reading ability. Kappa is a measure of agreement used with categorical data typically used to assess the degree to which two raters agree on a categorical decision. *K* is considered to be a more robust measure than basic percent agreement because it accounts for agreement occurring by chance. Therefore, it is considered to be a more conservative

measure than basic percent agreement. In the current context, $K$ was used to assess the extent to which each short-form and each long-form agree in labeling students as at-risk or advanced. The equation for $K$ is

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{16}$$

where Pr($a$) is the relative observed agreement among raters and Pr($e$) is the hypothetical probability of chance agreement. The observed data is used to calculate the probabilities of each observation occurring by chance for each category. If the short-form and long-form are in complete agreement, then $K = 1$. If there is no agreement between forms, then $K \leq 0$. Coefficients may range from -1.0 to +1.0 with scores less than 0 indicating no agreement, and scores ranging from 0 to .20 as slight, .21 to .40 as fair, .41 to .60 as moderate, .60 to .80 as substantial, and .81 to 1.0 considered almost perfect agreement (Landis & Koch, 1977).

Finally, the degree to which each short-form detected at-risk and advanced second grade readers with sensitivity and specificity was analyzed (Altman & Bland, 1994). The standard scores that delineate the top and bottom 20% of students in the second grade sample were selected as the cut scores. Next, five 2 x 2 frequency matrixes were created for each long-form resulting in a total of ten matrixes. A template of the matrix used to estimate sensitivity and specificity is illustrated in Table 2.

TABLE 2: Matrix for Estimating Sensitivity and Specificity

| | Long-Form | | |
|---|---|---|---|
| Short-Form | Poor | Good | Total |
| Poor | a | b | |
| Good | c | d | |
| Total | | | |

a = True positives
b = False positives/over referrals
c = False negatives/underreferral
d = True negatives

The numbers of individuals correctly identified by each short-form are represented in cells a and d. Cell a represents true positives and cell d represents true negatives. The numbers of individuals who were not correctly identified are represented in cells b and c. Cell b represents false positives (i.e. over-referrals) in which students are inaccurately identified as being either at-risk or advanced by the short-form contrary to results on the long-form. Cell c represents false negatives (i.e. under-referrals) in which students are inaccurately identified as not being either at-risk or advanced by the short-form contrary to results on the long-form.

Using these matrixes, the sensitivity index is calculated by dividing the number of true positives (cell a) by the sum of true positives and false negatives (cell a + cell c). The specificity index was calculated by dividing the number of true negatives (cell d) by the sum of true negatives and false positives (cell d + cell b).

# CHAPTER IV: RESULTS

The results for the dimensionality assessment, short form development, correlation coefficients, Cohen's Kappa, as well as the sensitivity and specificity of each form are discussed separately.

## *Dimensionality Assessment*

A principal components analysis was conducted on the responses from students that completed both forms of the subtest. In all, there were 801 cases without missing responses. The results showed that the item pool satisfied the unidimensionality assumption of dichotomously scored IRT models such that the first factor accounted for 26% of the total variance while the second factor accounted for 14% of the total variance. Since the first factor accounted for more than 20% of the test variance, the item bank met Reckase's (1979) criteria for acceptable calibration using unidimensional models. Thus, the item bank was deemed appropriate for Rasch analysis.

## *Item Parameter Estimates*

The responses from the 801 students that completed both forms of the subtest without missing responses were used to calibrate item and person parameters. Results revealed four students that either answered every item correctly (n = 2) or incorrectly (n = 2). After deleting these extreme cases, the person and item parameters were recalibrated. The item parameter estimates for the 80-item bank was calibrated with a sample of 797 students. The results are presented in Table 3.

TABLE 3: Item Parameter Estimates For 80 Reading Items

| Item Entry | Item Number | Difficuty Parameter | Item Entry | Item Number | Difficuty Parameter |
|---|---|---|---|---|---|
| 1 | a1 | -3.12 | 41 | b1 | -2.55 |
| 2 | a2 | -2.42 | 42 | b2 | -2.27 |
| 3 | a3 | -2.47 | 43 | b3 | -2.25 |
| 4 | a4 | -2.40 | 44 | b4 | -2.52 |
| 5 | a5 | -2.23 | 45 | b5 | -2.20 |
| 6 | a6 | -2.22 | 46 | b6 | -2.20 |
| 7 | a7 | -2.25 | 47 | b7 | -1.96 |
| 8 | a8 | -1.80 | 48 | b8 | -1.96 |
| 9 | a9 | -1.32 | 49 | b9 | -1.39 |
| 10 | a10 | -1.21 | 50 | b10 | -1.13 |
| 11 | a11 | -0.95 | 51 | b11 | -0.96 |
| 12 | a12 | -0.55 | 52 | b12 | -0.86 |
| 13 | a13 | -0.59 | 53 | b13 | -0.53 |
| 14 | a14 | -0.46 | 54 | b14 | -0.54 |
| 15 | a15 | -0.33 | 55 | b15 | -0.32 |
| 16 | a16 | -0.12 | 56 | b16 | -0.24 |
| 17 | a17 | -0.14 | 57 | b17 | -0.18 |
| 18 | a18 | -0.02 | 58 | b18 | 0.02 |
| 19 | a19 | 0.02 | 59 | b19 | 0.30 |
| 20 | a20 | 0.56 | 60 | b20 | 0.42 |
| 21 | a21 | 0.54 | 61 | b21 | 0.61 |
| 22 | a22 | 0.52 | 62 | b22 | 0.76 |
| 23 | a23 | 0.75 | 63 | b23 | 0.65 |
| 24 | a24 | 0.71 | 64 | b24 | 0.95 |
| 25 | a25 | 0.83 | 65 | b25 | 0.99 |
| 26 | a26 | 1.02 | 66 | b26 | 1.19 |
| 27 | a27 | 1.11 | 67 | b27 | 1.08 |
| 28 | a28 | 1.29 | 68 | b28 | 1.31 |
| 29 | a29 | 1.35 | 69 | b29 | 1.28 |
| 30 | a30 | 1.38 | 70 | b30 | 1.49 |
| 31 | a31 | 1.27 | 71 | b31 | 1.50 |
| 32 | a32 | 1.61 | 72 | b32 | 1.78 |
| 33 | a33 | 1.72 | 73 | b33 | 1.61 |
| 34 | a34 | 1.72 | 74 | b34 | 1.78 |
| 35 | a35 | 1.76 | 75 | b35 | 1.84 |
| 36 | a36 | 1.55 | 76 | b36 | 2.04 |
| 37 | a37 | 1.90 | 77 | b37 | 2.19 |
| 38 | a38 | 2.13 | 78 | b38 | 2.00 |
| 39 | a39 | 1.90 | 79 | b39 | 2.19 |
| 40 | a40 | 2.19 | 80 | b40 | 2.45 |

## *Model Fit Analysis*

Model fit analyses were conducted in conjunction with analyses of parameter invariance. According to Bond and Fox (2007), outfit or infit mean-squares greater than or equal to 2.0 can distort measurement. There were 18 items that failed to meet these criteria. Inspection of person fit statistics showed that 69 examinees also exceeded these criteria. When item parameters were calibrated without these examinees, only one item exceeded the criteria. According to Linacre (2010, personal communication), "Rasch models are remarkably robust against misfitting people. Often they make the fit statistics look bad, but have almost no influence on the estimated measures." In order to determine if this threat to item parameter invariance was inconsequential, the item parameters were calibrated with and without misfitting persons. The item difficulty estimates for the complete sample (n = 797) and subsample without significant misfits (n = 728) were plotted on the corresponding x and y axes in Figure 5. The plotted points lie closely along the center diagonal line and within the control lines for a 95% confidence band around the diagonal. This indicates that the misfitting persons are not distorting measurement. Therefore, the persons with outfit or infit mean-squares greater than or equal to 2.0 were included in the subsequent analyses.

The invariance of person parameters was examined by dividing the test in half based on item difficulty and calibrating person parameters using each half of the sample. The scatter plot is shown in Figure 6. Again, the plotted points lie closely along the center diagonal line and within the control lines for a 95% confidence band around the diagonal which demonstrates the invariance of the person parameters.

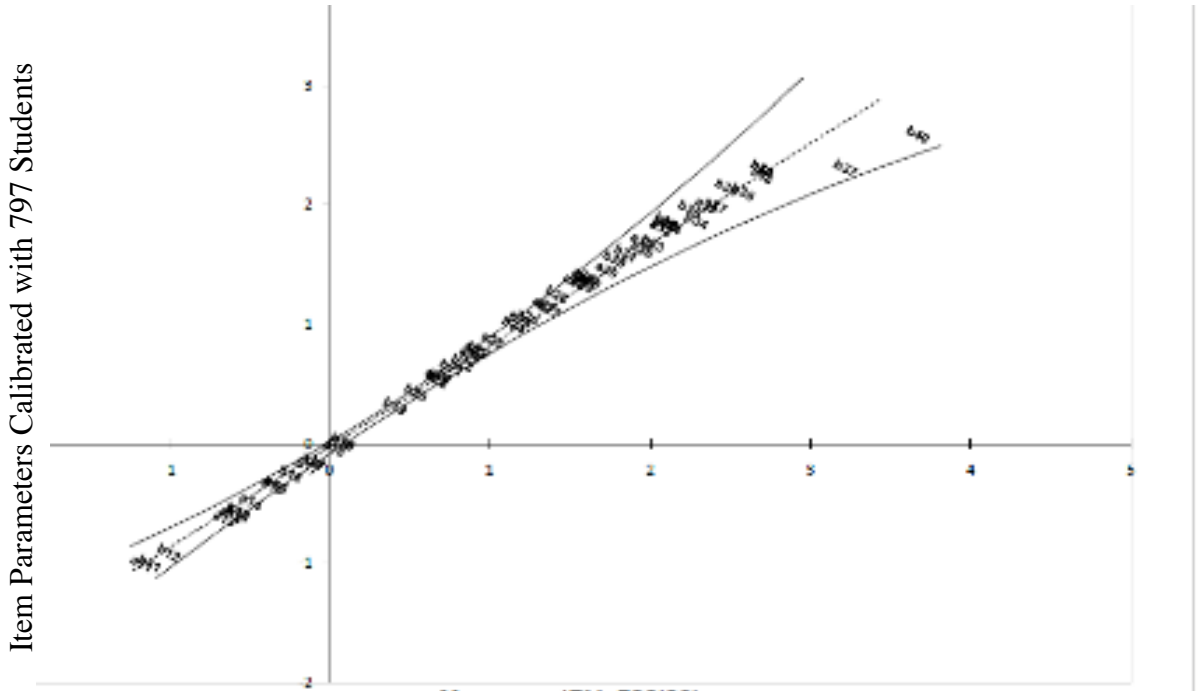


FIGURE 5: Item Parameter Invariance Assessment Using Item Parameter Calibration With and Without Misfitting Persons.

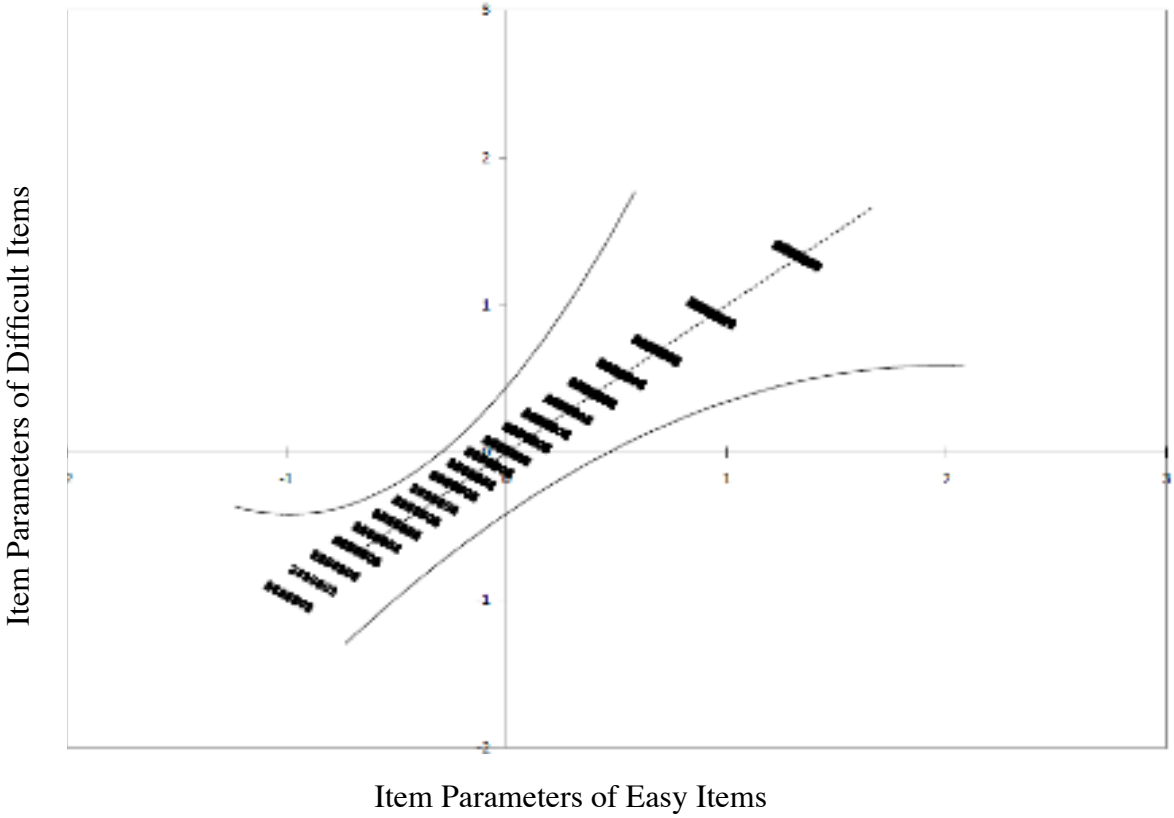


FIGURE 6: Person Parameter Invariance Assessment Using Difficult Versus Easy Items

Finally, the robustness of item estimates (n = 80) was assessed by dividing the sample of persons in half according to ability and calibrating item parameters for the total test using each half of the sample. The scatter plot is presented in Figure 7. Visual inspection revealed that 6 of 80 item locations fall outside of the control lines (i.e., 7.5%). These items were excluded from the final item bank. After deleting misfitting items, the item parameters were recalibrated. The parameter estimates for the 74 items included in the final item bank are provided in Appendix A. The parameter estimates for the 78 second grade students are provided in Appendix B.

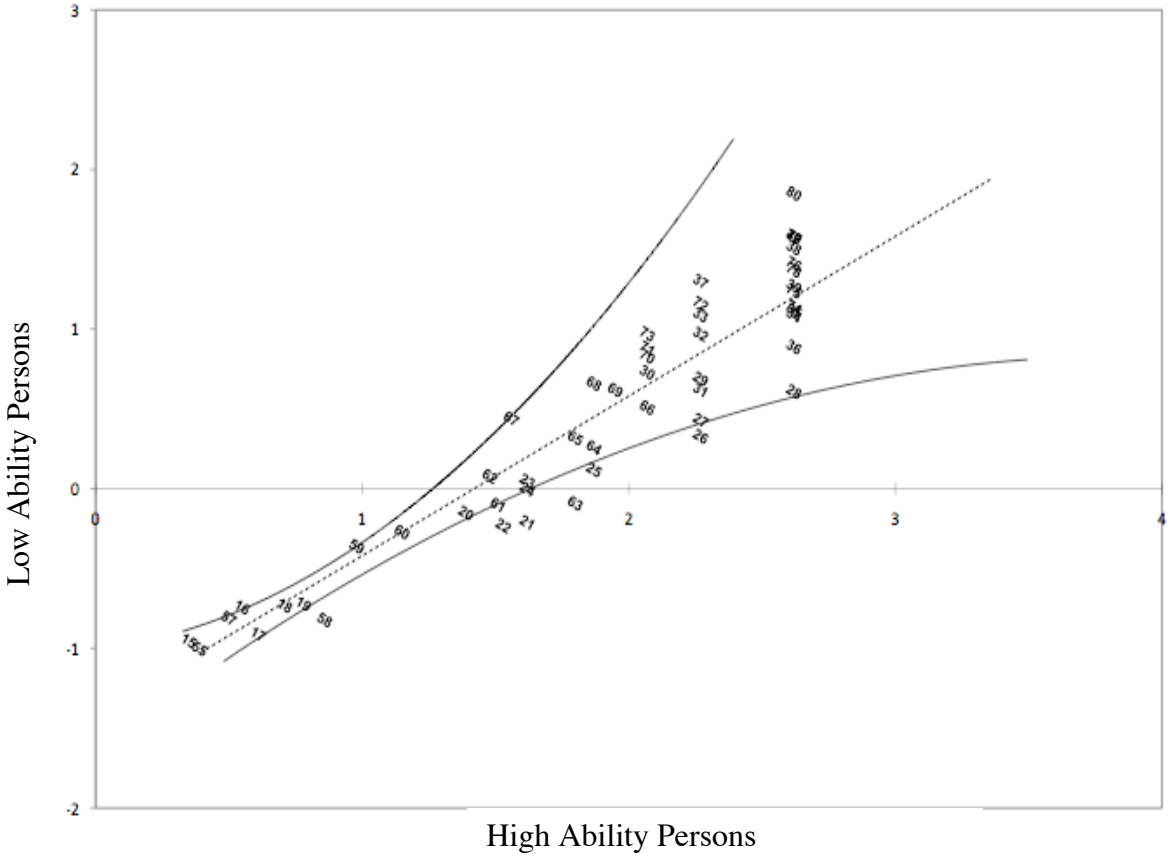


FIGURE 7: Item Parameter Invariance Assessment Using Low Ability vs. High Ability Persons

## *Short Form Development*

The ability estimates for the second grade students without missing data included in the normative sample (n = 78) were calculated with the Maximum Likelihood Estimation (MLE) method and ranked across the ability continuum. Two cut points were set to the θ values delineating the top 20% (θ = -0.14) and bottom 20% (θ = -2.34) of the sample resulting in three performance categories: At-Risk, On-Track, and Advanced.

Item information functions for the 74 items included in the final item bank were used to select items for short-forms to create an optimal test information function. Items that provided the most information closest to the θ value delineating the at-risk group and the advanced group were selected. This method was repeated iteratively until the predetermined number of items for each form (n = 10, 16, 22, 28, 34 items) was satisfied. On each form, an equal number of items were selected to target at-risk and advanced readers.

The test information functions for the five short forms and both long forms of reading items are presented in Figure 8. Test information indicates the precision of measurement on a scaled metric. It is derived as the expected value o the inverse of the squared standard error of measurement. The test information functions for the short-forms are symmetric and centered around an ability level of -1.25 while the test information functions for both long forms are centered around an ability level of +0.75. This indicates that the short forms provide better measurement precision for the target ability groups (i.e., beginning readers) than the long forms.

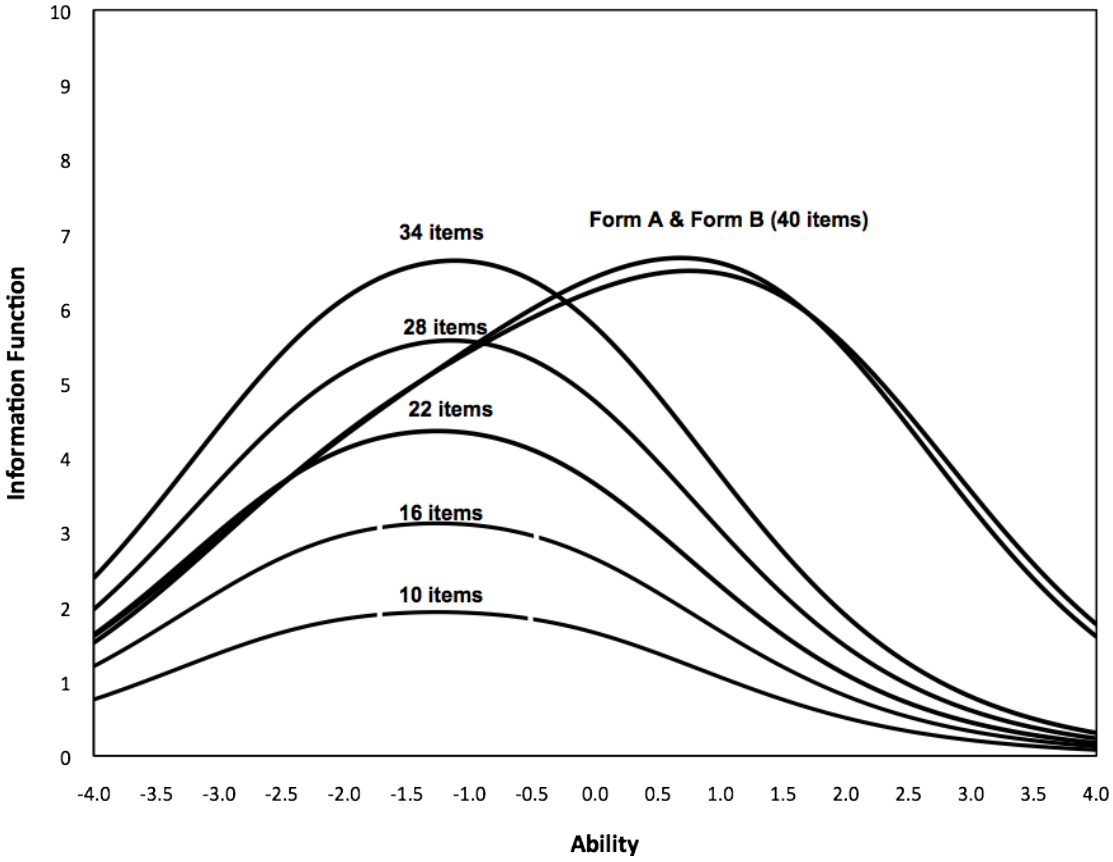


*FIGURE 8: Test Information for Reading Items Across Forms*

## *Correlation Coefficients and Classification Accuracy*

The bivariate correlation coefficients and classification accuracy of the short-forms as compared to the full-length measures are summarized in Table 4. The revised scales with 10 items (0.92, 0.95), 16 items (0.95, 0.97), 22 items (0.96, 0.98), 28 items (0.98, 0.99), and 34 items (0.98, 1.0) were strongly correlated with Form A and Form B, respectively.

TABLE 4: Correlation and Kappa Coefficients for Five Short Forms and Two Long Forms

| Test Forms | Correlation | | | At-Risk Group Kappa | | | Advanced Group Kappa | | |
|---|---|---|---|---|---|---|---|---|---|
| | Form A | Form B | Average | Form A | Form B | Average | Form A | Form B | Average |
| 10-Item Form | 0.92 | 0.95 | 0.94 | 0.63 | 0.67 | 0.65 | 0.79 | 0.86 | 0.83 |
| 16-Item Form | 0.95 | 0.97 | 0.96 | 0.61 | 0.72 | 0.67 | 0.89 | 0.81 | 0.85 |
| 22-Item Form | 0.96 | 0.98 | 0.97 | 0.70 | 0.74 | 0.72 | 0.85 | 0.85 | 0.85 |
| 28-Item Form | 0.98 | 0.99 | 0.98 | 0.82 | 0.71 | 0.76 | 0.81 | 0.89 | 0.85 |
| 34-Item Form | 0.98 | 1.00 | 0.99 | 0.77 | 0.73 | 0.75 | 0.71 | 0.89 | 0.80 |

Correlation: Correlation of ability estimates with full-bank scores
Kappa: Kappa statistic as measure of classification consistency

A different pattern of performance emerged when it came to the classification consistency estimated by Cohen's kappa. For the Advanced Group, the 16-item scale performed somewhat better for the Advanced Group (Cohen's $K$ = .89) compared to the others on Form A ($K$ ranged from 0.71 to 0.89). Meanwhile, the 28 and 34 item tests performed best on Form B ($K$ = 0.89), however there was less variation between forms ($K$ ranged from 0.81 to 0.89). On average, the 16-item scale was the most effective and efficient ($K$ = .89) for the Advanced Group. For the At-Risk Group, the 28-item scale

($K$ = .82) performed best compared to others on Form A ($K$ ranged from 0.61 to 0.82). The 22-item scale ($K$ = .74) performed best on Form B while little variation was exhibited across forms ($K$ ranged from 0.67 to 0.74). On average, the 22-item scale was most effective and efficient for the At-Risk Group.

To determine the practical value of the short-forms, the sensitivity and specificity of each form was examined. The acceptable levels of sensitivity and specificity vary by field and according to the intended purpose of the test. If the purpose of the test is to ensure that truly at-risk students are identified in the area of beginning reading, Jenkins (2003) suggests a rigorous sensitivity standard of 0.90 and a specificity standard close to 0.90. In each analysis, short-form scores were used to predict performance on both full-form scores. The results are presented in Table 5.

Table 5. Performance Comparisons of Five Short Forms and Two Long Forms

| Test Forms | | | 10 Item Form | 16 Item Form | 22 Item Form | 28 Item Form | 34 Item Form |
|---|---|---|---|---|---|---|---|
| At-Risk Group | Form A | Sensitivity | 1.00 | 0.81 | 0.81 | 0.94 | 0.84 |
| | | Specificity | 0.81 | 0.87 | 0.92 | 0.94 | 0.95 |
| | Form B | Sensitivity | 1.00 | 0.88 | 0.82 | 0.82 | 0.76 |
| | | Specificity | 0.82 | 0.90 | 0.94 | 0.92 | 0.95 |
| | Average | Sensitivity | 1.00 | 0.85 | 0.82 | 0.88 | 0.79 |
| | | Specificity | 0.82 | 0.89 | 0.93 | 0.93 | 0.95 |
| Advanced Group | Form A | Sensitivity | 0.94 | 0.88 | 0.88 | 0.82 | 0.82 |
| | | Specificity | 0.92 | 0.98 | 0.97 | 0.97 | 0.97 |
| | Form B | Sensitivity | 1.00 | 0.82 | 0.88 | 0.88 | 0.88 |
| | | Specificity | 0.94 | 0.97 | 0.97 | 0.98 | 0.98 |
| | Average | Sensitivity | 0.97 | 0.85 | 0.88 | 0.85 | 0.85 |
| | | Specificity | 0.93 | 0.98 | 0.97 | 0.98 | 0.98 |

Sensitivity: true-positives/(true-positives + false-negatives)
Specificity: true-negatives/(true-negatives + false-positives)

In terms of accuracy and time-efficiency, the findings were most robust for the shortest form. Across long-form comparisons, the 10-item form identified readers in the At-Risk Group with the most sensitivity (1.0, 1.0) while simultaneously detecting readers in the Advanced Group with the most sensitivity (.94, 1.0). Conversely, the 34-item form identified at-risk readers (.81, .76) and advanced readers (.82, .88) with less sensitivity. The specificity levels were similar for the advanced group across all forms, but the 10-item form was the least accurate for the at-risk group (.81, .82). If the most important goal of an early reading screening measure is to identify students that are truly at-risk without over-identifying those that are not at-risk, the 10-item scale seems to be the most effective and efficient form. If the goal is to further identify advanced readers with sensitivity and specificity so teachers can use the outcomes to establish instructional reading groups, the 10-item scale continues to be the most effective and efficient form.

# Chapter V: DISCUSSION

The discussion is divided into four sections. First, the research question listed in the problem statement is addressed based on the results of the study. Second, implications for applied practice are considered. Third, limitations of the present investigation are reviewed. And finally, directions for future research are discussed.

## *Research Question*

*What short-form length (n = 10, 16, 22, 28, 34 items) optimizes the time-efficiency and classification accuracy of a beginning reading screening measure as compared to both full-length forms (n = 40 items each) in terms of simultaneously identifying at-risk and advanced second grade readers?*

The success of academic failure prevention models, such as Response to Intervention (RtI), hinges on accurately identifying which children are at-risk for academic failure (Compton et al., 2010). One method of increasing the probability of identifying at-risk students is to adjust the cut scores of screening measures. Increasing the cut score improves the sensitivity of the measure, which means identifying more true-positives (i.e., children that truly have reading problems). Problematically, this approach also tends to increase the percentage of false-positives (i.e., children that do not truly have reading problems). The consequence is that intervention services are provided to students that do not truly need them, which may squander limited educational resources (Jenkins & O'Connor, 2000). Alternatively, decreasing the cut score improves the specificity of the measure, which means less true positives and less false positives will be identified. The consequence of this action is that intervention services are not provided to the students that truly need them. Deciding which error is most egregious is ultimately a value

judgment. To maximize the effectiveness of RtI models, screening measures must avoid false positives (Jenkins, Hudson, & Johnson, 2007) and yield a high percentage of true positives – with sensitivity approaching 100% (Compton et al., 2010). An alternative method of increasing the accuracy of screening measures is to use item response theory models to improve the quality of the measurement scale – this was the method explored in the present investigation.

The purpose of this study was to address the measurement gap in the literature regarding the application of item response theory models to improve screening measures for beginning readers by conducting a Rasch analysis of a national diagnostic reading assessment developed with classical test theory methods. The goal was to optimize measurement precision and thereby enhance the classification accuracy of a screening measure designed to identify exceptional second grade readers representing both ends of the ability distribution. This involved examining the dimensionality and invariance of the item bank (n = 80 items) before creating and evaluating five short forms that varied by length (n = 10, 16, 22, 28, 34 items).

Given the goal of developing a screening measure to simultaneously identify at-risk and advanced second grade readers in order to enhance RtI frameworks, several priorities were established in order to determine the *best* form in terms of optimizing classification accuracy. Given the long-term personal and economic consequences of academic failure, the first priority was to identify poor readers with sensitivity. The second priority was to ensure that advanced students were also identified with a high degree of sensitivity. And the final priority was to attain acceptable levels of specificity in order to minimize the short-term economic consequences of providing services to students that do not have exceptional instructional needs. Thus, the ultimate goal was to develop a measure

with at least 90% average sensitivity compared to both long forms with a specificity level close to 90% for at-risk and advanced readers.

Based on these priorities, the 10-item form with 5 items devoted to detecting at-risk students and 5 items devoted to identifying advanced students seemed to optimize measurement precision best. This form achieved perfect sensitivity (average = 100%) in the identification of at-risk readers, which means the screener correctly identified every student performing in the bottom 20% of the second grade sample. At the same time, the specificity level was acceptable (average = 82%), though admittedly not ideal. Meanwhile, the 10-item form resulted in the highest sensitivity (average = 97%) for advanced readers with an excellent level of specificity (average = 93%) as well.

The present investigation extends extant research related to the development and analysis of beginning reading screening measures in several ways. First, item response theory techniques were used (1) to analyze a diagnostic assessment developed with classical test theory methods and then (2) to adapt the assessment as a tool to screen for exceptional readers. Compared to the average sensitivity (49%) and specificity (86%) levels of second grade reading screening measures for at-risk readers as reviewed by Jenkins, Hudson, and Johnson (2007), the sensitivity (100%) and specificity (82%) of the 10-item form developed in the current study were more robust given a priority of identifying at-risk readers. The observed levels of sensitivity and specificity for the 10-item form were more similar to the sensitivity (91%) and specificity (85%) levels reported by Foorman et al. (1998) in their review of the screening measure included in the Texas Primary Reading Inventory (TPRI). It should be noted that the TPRI was the only second grade screener included in the Jenkins, Hudson, and Johnson (2007) review that utilized item response theory rather than classical test theory techniques. Beyond the precision of measurement in identifying at-risk readers, the added benefit of the 10-item form

developed in the current investigation compared to all other screening measures developed to date for second grade readers is that the form can be used to simultaneously identify advanced readers with very high levels of sensitivity and specificity.

## *Implications for Applied Practice*

At the end of the day, elementary school teachers cannot begin to address the widespread educational failure plaguing the public school system without screening measures capable of accurately and efficiently identifying students that will need differentiated instruction in order to improve literacy skills. As such, there is a tremendous need for more precise and time efficient tools that can identify exceptional beginning readers. Most reading diagnostic and screening scales rely on classical test theory while rarely taking advantage of item response theory analysis in the development process. The advantage of using item response theory models is the ability to develop more precise and time efficient measures that can be customized to the needs of practitioners, psychometricians, or researchers in the field. As illustrated in the current study, test developers can use these models to select the minimum number of items that must be administered to achieve the desired level of measurement precision for the ability levels targeted by the assessment with short-form scores on the same measurement scale as scores derived from examinees that take longer forms.

Elementary school teachers are faced with a difficult challenge when asked each year to identify students at-risk for failure in core subject areas and to also ensure that all students make academic progress. The challenge is twofold. First, general education teachers must typically administer, score, and often interpret individually administered screening measures while simultaneously managing roughly 25-30 other students for

several hours or days until the assessment process is completed. Yet, the practical value of these efforts is somewhat questionable given the inadequate accuracy of many of the most widely used screening measures used with beginning readers. Second, teachers must decide how to react to the data which invariably indicates that some students can't read well or at all, others are fairly average, and some are reading above grade level. The response to such variation in most schools has been to teach to the average student, to provide extra help to the struggling students when time permits, and to be thankful for the most advanced students because they make the teaching process easier. As stated by Lyon and colleagues (2001), "There is no doubt that, because of limitations in training, many general education and special education teachers are not prepared to address and respond to these individual differences in an informed manner" (p.269). Nevertheless, identifying exceptional learners is the prerequisite step to providing effective instruction. Presently, teachers simply do not have the tools to efficiently identify low performing or high performing beginning readers.

Federal mandates stipulate that poor readers must be identified and should receive remedial instruction to prevent long-term failure. However, no such safeguards exist for academically advanced students despite the fact that if they do not receive instruction matched to instructional need, it is unlikely they will ever realize their academic potential. Instead, they are more apt to become less advanced over time. Consequently, these students will likely become casualties of the statistical phenomenon referred to as regression toward the mean unless provided with literary enrichment opportunities outside of school.

The outcomes of the current investigation suggest that by using the screening measure with only 10 items, second grade teachers can simultaneously (1) satisfy federal screening mandates related to the identification of poor readers and (2) establish

instructional groups for at-risk, on-track, and advanced readers in order to ensure that all students have an equal the opportunity to develop literacy skills. In this way, the screening measure was designed to compliment RtI practices while expanding the current conceptualization of such frameworks. While RtI has already made a profound impact in special education toward the goals of better achievement and behavioral outcomes for students identified with learning disabilities as well as students at-risk for developing one (Fletcher, Coulter, Reschley, & Vaughn, 2004), it is well suited for and could be significantly improved by simultaneously addressing the needs of highly capable students. As noted by Tilly (2009), "The purpose of RtI is squarely improving results for students: All students. Indeed, RtI is not about special education, nor general education, nor talented and gifted, nor at-risk, nor migrant education . . . RtI is about Every Education" (p. 12).

## *Limitations*

There were several limitations in the current investigation. First, the correlation coefficients, classification accuracy, sensitivity, and specificity of the short forms are based on comparisons to long forms of the same assessment rather than comparisons to external measures. Consequentially, the positive outcomes reported in this study may overestimate the accuracy of the short forms. Because of this limitation, the results should be interpreted cautiously.

Several circumstances under which the study was conducted may explain why the short form with only 10 items had a higher average level of sensitivity and specificity than many of the longer forms. First, the item bank was limited to 74 items with difficulty estimates that ranged from -3.18 to +2.48. Meanwhile, the ability estimates of the second grade students ranged from -3.06 to +0.27. Of the 74 items in the final item bank, only 38

were within the general range of these ability estimates. In other words, these items would be most appropriate to measure the reading skills of the second grade sample. However, the goal of the current study was to *screen* for at-risk and advanced students rather than estimate their general reading ability. As such, items that provide the most information near the established cut points for the at-risk ($\theta$ = -2.22) and advanced ($\theta$ = -0.17) groups optimize measurement precision. Of the 38 items within the general range of second grade ability estimates, 15 were within 0.5 logits of the established cut points (At-Risk Group = 7 items; Advanced Group = 8 items) while only 11 items were within 0.25 logits of the established cut points (At-Risk Group = 6 items; Advanced Group = 5 items). Due to the limited item bank size, less informative items were necessarily selected to satisfy the requisite number of items needed for the longer forms. These circumstances presumably led to higher levels of sensitivity and specificity for the 10-item form, but lower levels for the initial short forms with more items. As the number of items on the short forms approximated the number of items on the long forms, classification accuracy increased. In general, the results show that the presence of items near the diagnostic thresholds in the sample is essential for improving the accuracy of screening measures. Given a larger item bank with more items located near the established cut point, sensitivity and specificity levels of the succession of short forms may have been improved.

An additional concern is that the second grade sample size was limited which may influence the stability of item parameters. Accordingly, the observed classification consistency results should not be generalized beyond the specific item bank and the selected cut score locations examined in this study. Moreover, the insufficient sample size prevented an analysis of the accuracy of the short forms to identify at-risk and advanced students in the beginning of the year compared to the accuracy of identification at the end of the year. Since many schools administer beginning reading screening measures in the

fall and spring of the academic school year, it is important to develop and analyze measures that are sensitive to developmental changes in reading ability at these times.

## *Directions for Future Research*

The current investigation could be extended in several ways. First, more research is needed to examine the utility of the short-forms collected in the beginning and end of the year in terms of the concurrent and predictive validity with external, high-stakes assessments. These results should then be cross-validated with other samples collected in the fall and spring to further examine the reliability and validity of estimates. This cross-validation research should be conducted periodically since demographic changes can affect the precision of prediction models. In the event a significant discrepancy is identified, the prediction model must be recalibrated using the most current school or district data in order to maximize the utility of the screening measures.

To optimize teacher time while reducing scoring errors, subsequent studies should also evaluate changes in measurement precision for short-forms administered using a computer. Computerized screening measures can be readily created using the item format involved in this study. One of the advantages of using such tests is that scoring is automatized which means teachers could immediately access a class report of assessment outcomes and then use that information to plan instruction the following day rather than grading a stack of tests, transferring student scores to a spreadsheet, calculating individual and group ability estimates, and then analyzing the data before being able to plan instruction. The short forms could be static or they could be administered adaptively based on the examinee's responses to previous questions. Since computer adaptive tests (CATs) maximize information about the individual's probable score, they typically require fewer

items than static short forms to attain comparable measurement precision (Choi, Reise, Pilkonis, & Cella, 2010).

Revolutionary research is also possible. For instance, assessments delivered via the computer may involve audio and video. They may constitute interactive simulations that make the presentation of information contingent on requests or actions. Test developers can also instantly measure more discrete behaviors and dimensions of behavior that may influence test performance. For example, response time to individual items may hold a key to better understanding and measuring latent abilities.

The frequency of behaviors through time which may be prerequisite skills of test mastery can also be instantly measured via the computer. Such data can be used to answer important empirical questions. For instance, is the fluency with which students' type their name or correctly answering certain types of items positively correlated with better outcomes?

Fluency commonly refers to the speed and accuracy or smoothness of a performance. As derived from the behavior analytic or biobehavioral process approach (after Donahoe & Palmer, 1994), fluency is not measured by reaction time, but as a count–per–minute in the free operant tradition of the early behavior analytic laboratory (Ferster & Skinner, 1957; Skinner, 1938). Free operant procedures arrange a situation "such that responses may be emitted freely, and the emission of each response leaves the organism free to initiate the next response" (Kling & Riggs, 1971, p. 602). Free operant procedures, in contrast to discrete trial procedures, provide the subject with unlimited opportunities to respond within each experimental session. Rate is typically used to characterize fluency observations because the metric is sensitive to both the speed and accuracy of performance. It is calculated by dividing the total number of observed behaviors by the total time spent recording, which results in estimates of behavior units per minute (e.g., 50

letters written correctly / 30 seconds = 100 letters written correct per minute) (Pennypacker, Koenig, & Lindsley, 1972). Accordingly, the concept of fluency is based upon observed frequencies of behavior and certain predictions that can be made from those frequencies. The predictions that can be made relative to the speed with which students can read grade level text, answer basic math fact problems, or write the letters of the alphabet in one-minute is an example of this principle.

A review of research analyzing the academic abilities of students at different developmental stages suggests that those with learning problems often have trouble mastering fundamental skills in math (Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008; Geary, 2003), reading (Chard, Vaughn, & Tyler, 2002; Pikulski & Chard, 2005) and writing (Brooks, Vaughan, & Berninger, 1999; Graham, Harris, & Fink, 2000). As such, the fluency metric can be used to uniquely discriminate between expert and novice performance. To illustrate this point, Fleischner, Garnett, and Shepherd (1982) compared the math fact computation skills of elementary school students identified as having learning disabilities with average students. They reported that performance was essentially indistinguishable based on the measure of percentage correct. However, on timed assessments, the students identified with learning disabilities completed only one-third as many math fact problems as their non-identified peers. In the area of reading, research shows that the combination of the slope and intercept of word reading fluency skills significantly improves the classification accuracy of a base screening battery for first grade students (Compton, Fuchs, Fuchs, & Bryant, 2006; Compton et al., 2010). On the basis of this research and other findings showing how basic skill fluency can uniquely distinguish between advanced and at-risk students, most universal screening and progress monitoring measures used within RtI frameworks are rate-based measures. Using item response theory models to analyze the interactions between discrete responses, basic skill

rates, and complex skill rates within a behavioral fluency framework will deepen understanding about the nature of the relationship between test behavior and latent ability.

Finally, most tests developed with item response theory techniques use items that are calibrated using unidimensional models, which assume that all items within an assessment measure the same construct. Multidimensional item response theory models also exist. Where the probability of successfully answering a test item depends on one underlying ability with unidimensional item response theory models, multidimensional models simultaneously take into account multiple basic abilities required to answer individual test items (Embretson & Reise, 2000). Depending on the extent to which correctly answering test items requires more than one skill, multidimensional models may allow more precise modeling of test behavior. For example, these models could be used to model the probability of solving a math word problem as a function of a combination of different abilities such reading ($\theta_1$) and math computation skills ($\theta_2$). Similarly, they may be useful in modeling the probability of correctly answering the items that require both word reading ($\theta_1$) and vocabulary skills ($\theta_2$) while factoring in oral reading fluency ($\theta_3$). Though limited, research indicates that compared to outcomes achieved using unidimensional item response theory models, substantial gains in measurement precision can be achieved by the appropriate application of multidimensional models (Segall, 1999). As such, it's possible that such models can significantly improve the precision of beginning reading screening measures. Future research should therefore examine the extent to which multidimensional models can be applied to the current dataset and recover a multi-dimensional structure. Given adequate fit, does the multidimensional model recover the true item and person parameters more accurately and efficiently than the unidimensional Rasch model? If so, could the slope of reading fluency measures collected over a short

time period function as an independent "item" and further improve the precision of the measurement model? The possible extensions for future research are seemingly endless.

In conclusion, if educational researchers, practitioners, and students are to derive the benefits of fundamental measurement necessary for the development of rational quantitative human sciences, future research must incorporate the principles and procedures of item response theory modeling. As noted by Johnston and Pennypacker (1993), perhaps the greatest difficulty of using indirect measures is accurately understanding the relationship between what is actually being measured by a set of test items and what the success or failure on those test items is supposed to represent. It is incumbent upon those that use indirect methods to provide scientific evidence that the measurement tools are valid and reliable measures of a given skill proficiency, especially in matters of education. In the end, the future of the learning sciences and the success of system-wide efforts like RtI to improve the quality of the public education system may reside in the intelligent application of item response models to the theory-practice nexus of empirical research.

## *APPENDIX A*

Item Parameter Estimates for 74 Reading Items

| Item Entry | Item Number | Difficuty Estimate | Item Entry | Item Number | Difficuty Estimate |
|---|---|---|---|---|---|
| 1 | a1 | -3.18 | 38 | b2 | -2.33 |
| 2 | a2 | -2.48 | 39 | b3 | -2.31 |
| 3 | a3 | -2.52 | 40 | b4 | -2.57 |
| 4 | a4 | -2.45 | 41 | b5 | -2.25 |
| 5 | a5 | -2.29 | 42 | b6 | -2.25 |
| 6 | a6 | -2.27 | 43 | b7 | -2.02 |
| 7 | a7 | -2.31 | 44 | b8 | -2.02 |
| 8 | a8 | -1.85 | 45 | b9 | -1.44 |
| 9 | a9 | -1.37 | 46 | b10 | -1.18 |
| 10 | a10 | -1.26 | 47 | b11 | -1.01 |
| 11 | a11 | -1.00 | 48 | b12 | -0.92 |
| 12 | a12 | -0.61 | 49 | b13 | -0.59 |
| 13 | a13 | -0.65 | 50 | b14 | -0.59 |
| 14 | a14 | -0.52 | 51 | b15 | -0.38 |
| 15 | a15 | -0.38 | 52 | b16 | -0.30 |
| 16 | a16 | -0.19 | 53 | b17 | -0.24 |
| 17 | a17 | -0.21 | 54 | b19 | 0.22 |
| 18 | a18 | -0.09 | 55 | b20 | 0.33 |
| 19 | a19 | -0.05 | 56 | b21 | 0.52 |
| 20 | a20 | 0.47 | 57 | b22 | 0.66 |
| 21 | a23 | 0.65 | 58 | b24 | 0.85 |
| 22 | a24 | 0.61 | 59 | b25 | 0.89 |
| 23 | a27 | 1.01 | 60 | b26 | 1.07 |
| 24 | a28 | 1.18 | 61 | b27 | 0.97 |
| 25 | a29 | 1.24 | 62 | b28 | 1.19 |
| 26 | a30 | 1.28 | 63 | b29 | 1.16 |
| 27 | a31 | 1.16 | 64 | b30 | 1.35 |
| 28 | a32 | 1.50 | 65 | b31 | 1.40 |
| 29 | a33 | 1.62 | 66 | b32 | 1.66 |
| 30 | a34 | 1.62 | 67 | b33 | 1.47 |
| 31 | a35 | 1.66 | 68 | b34 | 1.68 |
| 32 | a36 | 1.43 | 69 | b35 | 1.76 |
| 33 | a37 | 1.81 | 70 | b36 | 1.92 |
| 34 | a38 | 2.02 | 71 | b37 | 2.08 |
| 35 | a39 | 1.76 | 72 | b38 | 1.88 |
| 36 | a40 | 2.08 | 73 | b39 | 2.08 |
| 37 | b1 | -2.60 | 74 | b40 | 2.48 |

## *APPENDIX B*

Person Parameter Estimates For 78 Second Grade

Students Based on Responses to 74 Items

| N | Ability Parameter | N | Ability Parameter |
|---|---|---|---|
| 1 | -3.06 | 40 | -1.45 |
| 2 | -3.06 | 41 | -1.34 |
| 3 | -2.77 | 42 | -1.34 |
| 4 | -2.77 | 43 | -1.34 |
| 5 | -2.77 | 44 | -1.24 |
| 6 | -2.67 | 45 | -1.14 |
| 7 | -2.67 | 46 | -0.95 |
| 8 | -2.67 | 47 | -0.95 |
| 9 | -2.58 | 48 | -0.78 |
| 10 | -2.58 | 49 | -0.78 |
| 11 | -2.49 | 50 | -0.78 |
| 12 | -2.49 | 51 | -0.78 |
| 13 | -2.41 | 52 | -0.70 |
| 14 | -2.41 | 53 | -0.70 |
| 15 | -2.34 | 54 | -0.63 |
| 16 | -2.26 | 55 | -0.56 |
| 17 | -2.26 | 56 | -0.56 |
| 18 | -2.26 | 57 | -0.49 |
| 19 | -2.26 | 58 | -0.42 |
| 20 | -2.26 | 59 | -0.35 |
| 21 | -2.18 | 60 | -0.35 |
| 22 | -2.18 | 61 | -0.28 |
| 23 | -2.18 | 62 | -0.21 |
| 24 | -2.18 | 63 | -0.21 |
| 25 | -2.18 | 64 | -0.21 |
| 26 | -2.09 | 65 | -0.21 |
| 27 | -2.09 | 66 | -0.21 |
| 28 | -2.09 | 67 | -0.21 |
| 29 | -2.09 | 68 | -0.21 |
| 30 | -2.09 | 69 | -0.14 |
| 31 | -2.00 | 70 | -0.07 |
| 32 | -2.00 | 71 | 0.00 |
| 33 | -1.90 | 72 | 0.00 |
| 34 | -1.69 | 73 | 0.00 |
| 35 | -1.69 | 74 | 0.07 |
| 36 | -1.69 | 75 | 0.13 |
| 37 | -1.57 | 76 | 0.20 |
| 38 | -1.45 | 77 | 0.20 |
| 39 | -1.45 | 78 | 0.27 |

## *APPENDIX C*

Person Parameter Estimates Across

Forms for 78 Second Grade Students

| N | Form A 40 Items | Form B 40 Items | 10 Items | 16 Items | 22 Items | 28 Items | 34 Items |
|---|---|---|---|---|---|---|---|
| 1 | -3.17 | -3.06 | -3.09 | -2.86 | -2.94 | -3.00 | -3.10 |
| 2 | -3.61 | -3.06 | -2.72 | -2.86 | -2.94 | -3.00 | -2.88 |
| 3 | -2.64 | -2.77 | -3.09 | -2.86 | -2.72 | -2.78 | -2.74 |
| 4 | -3.17 | -2.77 | -2.72 | -2.62 | -2.72 | -2.78 | -2.88 |
| 5 | -2.86 | -2.77 | -2.44 | -2.62 | -2.57 | -2.51 | -2.63 |
| 6 | -2.31 | -2.67 | -2.44 | -2.46 | -2.44 | -2.51 | -2.63 |
| 7 | -2.64 | -2.67 | -2.72 | -2.62 | -2.57 | -2.63 | -2.63 |
| 8 | -2.47 | -2.67 | -2.72 | -2.86 | -2.72 | -2.63 | -2.63 |
| 9 | -2.86 | -2.58 | -2.72 | -2.62 | -2.72 | -2.63 | -2.63 |
| 10 | -2.64 | -2.58 | -2.22 | -2.17 | -2.23 | -2.31 | -2.44 |
| 11 | -2.13 | -2.50 | -2.72 | -2.46 | -2.33 | -2.41 | -2.44 |
| 12 | -2.86 | -2.50 | -2.22 | -2.46 | -2.44 | -2.41 | -2.53 |
| 13 | -2.64 | -2.42 | -2.22 | -2.32 | -2.33 | -2.41 | -2.36 |
| 14 | -2.64 | -2.42 | -2.22 | -2.32 | -2.33 | -2.41 | -2.44 |
| 15 | -2.13 | -2.34 | -2.22 | -2.32 | -2.33 | -2.31 | -2.36 |
| 16 | -2.64 | -2.26 | -2.22 | -2.17 | -2.23 | -2.31 | -2.27 |
| 17 | -2.31 | -2.26 | -1.95 | -2.17 | -2.23 | -2.21 | -2.27 |
| 18 | -2.31 | -2.26 | -2.44 | -2.46 | -2.33 | -2.31 | -2.27 |
| 19 | -2.31 | -2.26 | -1.95 | -2.17 | -2.23 | -2.21 | -2.27 |
| 20 | -1.93 | -2.26 | -2.22 | -2.32 | -2.33 | -2.31 | -2.27 |
| 21 | -2.47 | -2.18 | -2.22 | -2.17 | -2.11 | -2.11 | -2.18 |
| 22 | -2.31 | -2.18 | -2.22 | -2.32 | -2.23 | -2.21 | -2.18 |
| 23 | -2.31 | -2.18 | -2.22 | -2.32 | -2.23 | -2.21 | -2.18 |
| 24 | -1.93 | -2.18 | -2.22 | -2.32 | -2.23 | -2.21 | -2.18 |
| 25 | -1.93 | -2.18 | -2.22 | -2.32 | -2.44 | -2.21 | -2.18 |
| 26 | -2.13 | -2.09 | -2.22 | -2.17 | -1.99 | -2.01 | -2.09 |
| 27 | -1.93 | -2.09 | -1.31 | -2.01 | -2.11 | -2.11 | -2.09 |
| 28 | -2.13 | -2.09 | -1.31 | -1.78 | -2.11 | -2.11 | -2.09 |
| 29 | -2.13 | -2.09 | -2.22 | -2.01 | -2.11 | -2.11 | -2.09 |
| 30 | -1.69 | -2.09 | -2.22 | -2.17 | -2.23 | -2.01 | -2.00 |
| 31 | -2.31 | -2.00 | -1.95 | -2.01 | -1.99 | -2.01 | -2.00 |
| 32 | -2.13 | -2.00 | -2.22 | -2.01 | -1.99 | -2.01 | -2.00 |
| 33 | -1.93 | -1.91 | -1.95 | -1.78 | -1.84 | -1.89 | -1.89 |
| 34 | -1.69 | -1.69 | -1.31 | -1.34 | -1.33 | -1.59 | -1.63 |
| 35 | -1.93 | -1.69 | -1.95 | -2.01 | -1.84 | -1.59 | -1.63 |
| 36 | -1.69 | -1.69 | -1.31 | -1.34 | -1.33 | -1.59 | -1.63 |
| 37 | -1.69 | -1.57 | -1.31 | -1.34 | -1.64 | -1.40 | -1.48 |
| 38 | -1.45 | -1.46 | -1.31 | -1.34 | -1.33 | -1.19 | -1.32 |
| 39 | -1.45 | -1.46 | -1.31 | -1.34 | -1.33 | -1.19 | -1.32 |
| 40 | -1.45 | -1.46 | -1.95 | -1.78 | -1.64 | -1.40 | -1.32 |
| 41 | -1.45 | -1.35 | -1.31 | -1.34 | -1.33 | -1.19 | -1.32 |
| 42 | -1.22 | -1.35 | -1.31 | -1.34 | -1.33 | -1.19 | -1.16 |
| 43 | -1.22 | -1.35 | -1.31 | -1.34 | -1.33 | -1.19 | -1.32 |
| 44 | -1.45 | -1.24 | -1.31 | -0.89 | -1.02 | -1.01 | -1.16 |
| 45 | -1.02 | -1.14 | -1.31 | -1.34 | -1.02 | -1.01 | -1.01 |
| 46 | -1.02 | -0.95 | -0.65 | -0.89 | -1.02 | -0.72 | -0.88 |
| 47 | -0.84 | -0.95 | -1.31 | -1.34 | -1.33 | -1.19 | -1.01 |
| 48 | -0.69 | -0.78 | -1.31 | -1.34 | -1.33 | -1.01 | -0.88 |
| 49 | -0.84 | -0.78 | -1.31 | -1.34 | -1.02 | -0.85 | -0.88 |
| 50 | -0.84 | -0.78 | -1.31 | -0.89 | -0.82 | -0.85 | -0.88 |
| 51 | -0.69 | -0.78 | -1.31 | -1.34 | -1.33 | -1.01 | -0.88 |
| 52 | -0.54 | -0.71 | -0.65 | -0.89 | -1.02 | -0.85 | -0.67 |
| 53 | -0.69 | -0.71 | -1.31 | -1.34 | -0.82 | -0.72 | -0.67 |
| 54 | -0.54 | -0.63 | -0.65 | -0.66 | -0.67 | -0.72 | -0.67 |
| 55 | -0.54 | -0.56 | -1.31 | -0.89 | -0.67 | -0.61 | -0.57 |
| 56 | -0.69 | -0.56 | -0.65 | -0.49 | -0.55 | -0.61 | -0.57 |
| 57 | -0.54 | -0.49 | -0.65 | -0.66 | -0.55 | -0.51 | -0.49 |
| 58 | -0.41 | -0.42 | -0.65 | -0.35 | -0.32 | -0.41 | -0.40 |
| 59 | -0.28 | -0.35 | -0.11 | 0.21 | -0.21 | -0.32 | -0.31 |
| 60 | -0.02 | -0.35 | -1.31 | -0.66 | -0.43 | -0.32 | -0.31 |
| 61 | -0.28 | -0.28 | -0.65 | -0.20 | -0.21 | -0.22 | -0.22 |
| 62 | -0.41 | -0.21 | -0.36 | -0.35 | -0.32 | -0.22 | -0.13 |
| 63 | -0.54 | -0.21 | -0.36 | -0.35 | -0.21 | -0.32 | -0.22 |
| 64 | -0.02 | -0.21 | -0.36 | -0.20 | -0.08 | -0.12 | -0.13 |
| 65 | -0.15 | -0.21 | -0.36 | -0.35 | -0.43 | -0.61 | -0.49 |
| 66 | -0.28 | -0.21 | -0.11 | -0.03 | -0.08 | -0.12 | -0.13 |
| 67 | -0.15 | -0.21 | -0.11 | -0.03 | -0.08 | -0.12 | -0.13 |
| 68 | -0.41 | -0.21 | -0.11 | -0.03 | -0.21 | -0.12 | -0.22 |
| 69 | -0.02 | -0.14 | 0.20 | -0.20 | -0.08 | 0.00 | -0.03 |
| 70 | -0.15 | -0.07 | -0.11 | -0.03 | 0.07 | 0.00 | 0.07 |
| 71 | -0.02 | 0.00 | 0.61 | 0.21 | 0.30 | 0.14 | 0.07 |
| 72 | -0.02 | 0.00 | -0.11 | -0.20 | -0.08 | 0.14 | 0.19 |
| 73 | 0.11 | 0.00 | 0.20 | 0.57 | 0.65 | 0.31 | 0.19 |
| 74 | -0.02 | 0.07 | 0.20 | 0.21 | 0.30 | 0.14 | 0.19 |
| 75 | -0.02 | 0.14 | 0.61 | 0.57 | 0.65 | 0.31 | 0.33 |
| 76 | -0.02 | 0.20 | 0.20 | 0.57 | 0.30 | 0.14 | 0.19 |
| 77 | 0.11 | 0.20 | 0.61 | 0.21 | 0.30 | 0.31 | 0.33 |
| 78 | 0.24 | 0.27 | 0.61 | 0.21 | 0.30 | 0.57 | 0.51 |

# References


Anastasia, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal, 308(6943),* 1552.

Barton, P., & Jenkins, L. (1995). *Literacy and dependency*. Princeton: Educational Testing Service [Policy Information Center].

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum.

Brooks, A., Vaughan, K., & Berininger, V. (1999). Tutorial interventions for writing disabilities: Comparison of transcription and text generation processes. *Learning Disability Quarterly, 22*, 183-190.

Brown, E. F., & Abernathy, S. H. (2009). Policy implications at the state and district level with RtI for gifted students. *Gifted Child Today, 32(3*), 52-57.

Bryant D. P., Bryant, Gersten, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008). Effects of tier 2 Intervention delivered as booster lessons mathematics interventions for first- and second-grade students with mathematics difficulties: The

effects of tier 2 intervention delivered as booster lessons. *Remedial and Special Education, 29,* 20-32.

Cattell, R. B. (1966). A scree test for the number of factors. *Multivariate Behavioral Research, 1(2),* 245-276.

Catts, H. W., Fey, M. E., Zhang X., & Tomblin J. B. (2001). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*, 331–361.

Catts H. W., Fey M. E., Zhang X., Tomblin J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implication. *Language, Speech, and Hearing Services in Schools, 32*, 38–50.

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics for Early Childhood Special Education, 12(2),* 196-211.

Chard, D., Vaughn, S., & Tyler, B. (2002). A Synthesis of Research on Effective Interventions for Building Reading Fluency with Elementary. *Journal of Learning Disabilities*, *35*, 386 - 406.

Choi, S., Reise, S., Pilkonis, P., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19(1),* 125-136.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Coleman, M. R., & Hughes, C. E. (2009). Meeting the needs of gifted students within an RtI framework. *Gifted Child Today, 32(3),* 14-17.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*, 327-340.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394–409.

Congress, U. S. (2001). No child left behind act of 2001. Public Law, 107–110.

Council of State Directors of Programs for the Gifted (1994). *The 1994 State of the States Gifted and Talented Education Report.* Austin, TX: Author.

Dodd, B. G. (1985). *Attitude scaling: A comparison of the graded response and partial credit latent trait models*. Unpublished doctoral dissertation, University of Texas at Austin.

Donahoe, J., & Palmer, D. (1994). *Learning and complex behavior*. Needham Heights, MA: Allyn and Bacon.

Drasgow, F. & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of industrial and organizational psychology: Vol. 1* (2nd ed., p. 577-636). Palo Alto, CA: Consulting Psychologists Press.

Dunn, L. M. & Dunn, L. M. (1981). *Peabody picture vocabulary test-Revised*. Circle Pines, MN: American Guidance Service.

Embretson, S. E., and Reise, S. P. (2000*) Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ferster, C. B. & Skinner, B. F. (1957). *Schedules of reinforcement*. Englewood Cliffs, NJ: Prentice–Hall, Inc.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory of the general and academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality, 29*, 168–188.

Fleischner, J. E., Garnett, K., & Shepherd, M. J. (1982). Proficiency in basic fact computation of learning disabled and nondisabled children. *Focus on Learning Problems in Mathematics, 4,* 47-55.

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia, 54,* 304–331.

Foorman, B., & Ciancio, D. (2005). Screening for secondary intervention: Concept and context. *Journal of Learning Disabilities, 38*, 494– 499.

Foorman, B. R., Fletcher, J. M., Francis, D. J., Carlson, C. D., Chen, D., & Mouzaki, A., et al. (1998). *Technical Report: Texas Primary Reading Inventory* (1998 edition). Houston: Center for Academic and Reading Skills and University of Houston.

Fuchs, L. S., & Fuchs, D. (2007). Progress monitoring within a multi-tiered prevention system. *Perspectives, 33(2),* 43–47.

Geary, D. (2003). Learning disabilities in arithmetic: Problem-solving differences and cognitive deficits. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of*

*Learning Disabilities* (pp. 199-212). New York: Guilford.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*, 293–304.

Gellert, A., & Elbro, C. (1999). Reading disabilities, behaviour problems and delinquency: A review. *Scandinavian Journal of Educational Research, 43(2),* 131–155.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills*. Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility o a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257-288.

Graham, S., Harris, K., & Fink, B. (2000). Is handwriting causally related to learning to write? Treatment of handwriting problems in beginning writers. *Journal of Educational Psychology*, *92*, 620-633.

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools, 34*, 99-106.

Gredler, G. R. (2000). Early childhood screening for developmental and educational problems. In B.A. Bracken (Ed.*), The psychoeducational assessment of preschool children* (pp.399–411). Boston: Allyn & Bacon.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Harcourt Educational Measurement. (2002). *Stanford achievement test [SAT-10]*. San Antonio, TX.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogenous grouping in kindergarten: An application of marginal mean weighting through stratification. *Education Evaluation and Policy Analysis, 31(1),* 54-81.

The Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, §632, 118 Stat. 2744.

Jenkins, J. R. (2003). Candidate measures for screening at-risk students. Presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MD. Retrieved from http//:www.nrcld.org/symposium2003/jenkins/index.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for At-Risk Readers in a Response to Intervention Framework. *School Psychology Review, 36,* 582-600.

Jenkins, J. R., & O'Connor, R. (2000). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities* (pp. 99–149). Hillsdale, NJ: Erlbaum.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments*? Learning Disabilities Research & Practice, 24(4)*,174–185.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research*. Hillsdale, NJ: Erlbaum.

Kingslake, B. (2007). The predictive (in) accuracy of on-entry to school screening procedures when used to anticipate learning difficulties. *British Journal of Special Education, 10(4),* 23–26.

Kline, T. J. (2005). Psychological testing: A practical approach to design and evaluation. Thousand Oaks, CA: Sage.

Kling, J. W., & Riggs, L. A. (1971). *Woodworth & Schlosberg's Experimental Psychology,* 3rd Ed. New York: Holt, Rinehart and Winston, Inc.

Kurns, S., & Tilly D. (2008). *Response to Intervention blueprints for implementation: School building level edition*. NASDE, Alexandria, VA.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33(1*), 159-174.

Linacre, J. M. (2010). Winsteps (Version 3.65.0) [Computer Software]. Chicago: www.Winsteps.com.

Linacre, J. M. (2003). Constructing scientific measurement models. *Rasch Measurement Transactions*. *17(1),* 907.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, 7*.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of*

*Educational Measurement, 14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, C. L. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B. Schulte, A., & Olson, R. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.*), Rethinking special education for a new century* (pp.259–287). Washington, DC: Thomas B. Fordham Foundation.

McBee, M. T. (2006). A descriptive analysis of referral sources for gifted identification screening by race and socioeconomic status. *Journal of Secondary Gifted Education, 17(2),* 103–111.

McCardle, P., Scarborough H. S., & Catts H. W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning Disabilities Research & Practice, 16,* 230–239.

Messick, S. (1989). Validity. *Educational Measurement, 3(1)*, 13–103.

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425.

O'Connor, R. E., & Jenkins, J. R. (1999). The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3,* 159–197.

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. (1972). *Handbook of the standard celeration chart*. Kansas City, MO: Precision Media.

Pirani-McGurl, C. A. (2009). *The use of item response theory in developing a Phonics Diagnostic Inventory*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Pikulski, J., & Chard, D. (2005). Fluency: Bridge between decoding and reading comprehension. *Reading Teacher*, *58*, 510-519.

President's Commission on Excellence in Special Education. (2002). A new era: Revitalizing special education for children and their families. Retrieved from http://www.ed.gov/inits/commissionsboards

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reckase, M. D. (1979) Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4,* 207–230.

Reidel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42,* 546-567.

Ripley, A. (2010, December). Your child left behind. *Atlantic Monthly*, 94-98.

Samuels, S. J. (2007). Commentary: The DIBELS test is speed of barking at print: What we mean by reading fluency. *Reading Research Quarterly*, *42*, 563–566.

Sanders, W. (1999, September 1). Teachers, teachers, teachers. *DLC Blueprint Magazine*. Retrieved from http://www.dlc.org/ndol_ci.cfm?contentid=1199&kaid=110&subid=135

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007). Mini-mental status examination: a short form of MMSE was as accurate as the original MMSE in predicting dementia. *Journal of Clinical Epidemiology, 60,* 260–267.

Segall, D. O. (1999). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66(1)*, 79-97.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* Englewood Cliffs, NJ: Prentice–Hall.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, D.C.: National Academies Press.

State of the States Gifted and Talented Education Report (1994). Austin TX: Council of State Directors of Programs for the Gifted.

Steinberg, L. & Thissen, D. (1995). Item response theory methods in personality research. In P. E. Shrout & S. T. Fiske (Eds.) *Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stenner, A. J., & Horabin, I. (1992). Three stages of construct definition. *Rasch Measurement Transactions, 6*, 229.

Texas Education Agency. (2004). TPRI. Austin, TX: University of Texas System.

Tilly, D. (2009). Questions and answers on response to intervention. *Journal of Special Education Leadership, 50(4),* 7-12.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. A. (2007). Multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*, 225–256.

Vaughn, S., & Fuchs, L. S. (2003). Redefining LD as inadequate response to instruction: The promise and potential problems [Special issue]. *Learning Disabilities Research & Practice, 18(3)*, 137–146.

Vaughn, S., & Linan-Thompson, S. (2003). What is special about special education for students with learning disabilities*? Journal of Special Education, 37(3),* 140-147.

Walberg, H. J. (2001). Achievement in American schools. In T. M. Moe (Ed.), *American education: A primer* (pp. 43-68). Stanford: Hoover Institution Press.

Wechsler, D. (1972). *Wechsler Intelligence Scale for Children-Revised*. Cleveland, OH: The Psychological Corporation.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson psycho-educational battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson psychoeducational battery-3rd edition*. Itasca, IL: Riverside Publishing.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range, Inc. Retrieved from http://www.rasch.org/measess/me-all.pdf.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing problems* (pp.85-101). Princeton, N.J.: Educational Testing Service.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11(1)*, 57-67.

# VITA

Amy Broward Weisenburgh was born in Tampa, Florida on April 13, 1976, the daughter of Louis and Mary Weisenburgh. After graduating from Mercer Island High School located on Mercer Island, Washington in 1994, she began undergraduate study in Behavioral Analysis at the University of Florida. She received the degree of Bachelor of Science from the University of Florida in May 1999. In November of 1999, Amy began a Fulbright Research Grant in Comparative Education to design, implement, and evaluate a fluency-based math computation program at a school in Lima, Peru serving students with learning and behavior problems. After returning to the United States, Amy was employed as a teacher at Morningside Academy in Seattle, Washington. In 2002, she entered the Graduate School at the University of Washington. In 2003, she was employed as a special education teacher in Columbia, Missouri. She received the degree of Master of Arts in Special Education from The University of Washington and co-founded the non-profit organization Partnerships for Educational Excellence and Research (PEER) International in 2004.